The Bell System Technical Journal

July, 1934

The Compandor—An Aid Against Static in Radio Telephony *

By R. C. MATHES and S. B. WRIGHT

One of the important conditions which must be met by any speech transmission system is that it should transmit properly a sufficient range of speech intensities. In long-wave radio telephony, even after the speech waves are raised to the maximum intensity before transmission, there remain energy variations such that weak syllables and important parts of strong syllables may be submerged under heavy static. The compandor is an automatic device which compresses the range of useful signal energy variations at the transmitting end and expands the range to normal at the receiving end, thus improving the speech-to-noise ratio.

This paper deals with some of the fundamental characteristics of speech waves and explains how the task of changing them for transmission over the circuit and restoring them at the receiving end is accomplished. It is also shown that raising the strength of the weaker parts of speech gives these results: 1, the successful transmission of messages for a large percentage of the time previously uncommercial; 2, a reduction of the noise impairment of transmission for moderate and heavy static during time classed commercial; and 3, the ability to deliver higher received volumes due to the improved operation of the voice controlled switching circuits. In addition to these advantages, the compandor makes it possible to economize on radio transmitter power in times of light static.

Introduction

WHEN the original New York–London long-wave radiotelephone circuit was designed, it was recognized that radio noise would often limit transmission, especially for the weaker voice waves. Accordingly provision was made for manually adjusting the magnitude of the speech waves entering the radio transmitters to such a value as to load these transmitters to capacity.[1] While this treatment was very effective in improving the average speech-to-noise ratio and in preventing the strong peaks of speech from overloading the transmitter, it was, of course, unsuitable for following the rapidly varying amplitudes of the various speech sounds.

The total range of significant intensities applied to the circuit is in the order of 70 db, an energy ratio of 10 million to one. The manual adjustments referred to above were succesful in reducing this range to about 30 db. To further reduce this residual range an interesting

* Presented at Summer Convention of A. I. E. E., June, 1934. Published in *Electrical Engineering*, June, 1934.

device called the compandor has been developed. This device which works automatically makes a further reduction of one-half in the residual db range so that the range transmitted over the circuit is then only 15 db, an energy ratio of about 32 to one.

Speech Energy

Quantitative designation of speech intensity and hence of a range of intensities is rendered difficult by the rapidly varying amplitude characteristics of the various speech sounds. Devices called volume indicators are used fairly extensively to indicate the so-called "electrical volume" * of speech waves. A volume indicator is essentially a rectifier combined with a damped d-c. indicating meter on which are read in a specified manner the standard ballistic throws due to partly averaged syllables at a particular speech intensity. These devices are so designed and adjusted that they are insensitive to extremely high peak voltages of short duration, but their maximum deflection is approximately proportional to the mean power in the syllable. It has been found that, if commercial telephone instruments are used, the ear does not detect amplifier overloading of the extremely high peaks of short duration. Consequently, the volume indicator is a useful device for indicating the noticeable repeater overloading effect of a voice wave. These devices do not tell us much about the effect of the weaker voltages in overriding interference or operating voice-operated devices but they give a fairly satisfactory indication of loudness and possibilities of interference into other circuits.

The sound energy that the telephone transmits consists of complicated waves made up of tones of different pitch and amplitude. The local lines and trunks connecting the telephone to the subscribers' toll switchboard have little effect in changing the fundamental characteristics of these waves but, on account of various amounts of dissipation, the waves received at the toll switchboard are always weaker than those transmitted by the telephone. Furthermore, the strength of signals varies with the method of using the telephone, loudness of talking, battery supply, and transmitter efficiency. The subscriber may be talking over a long distance circuit from a distant city, in which case the loss of the toll line further attenuates the received waves. Figure 1 † shows that the range of outgoing speech volumes as measured by a volume indicator at the transatlantic switchboard at New York is nearly 40 db for terminal calls. When via calls and variation in volume

* The term volume will be used through the rest of this paper to designate this quantity and not as synonymous with loudness.[2]

† This curve is plotted on so-called *probability paper*, in which the scale is such that data distributed in accordance with the *normal law* will produce a straight line.

of the individual talker are taken into account, it is even greater than 40 db.

Volume Range of a Telephone Circuit

There are two limits on the range of volumes which a system can transmit. The upper limit of volume is set by the point at which

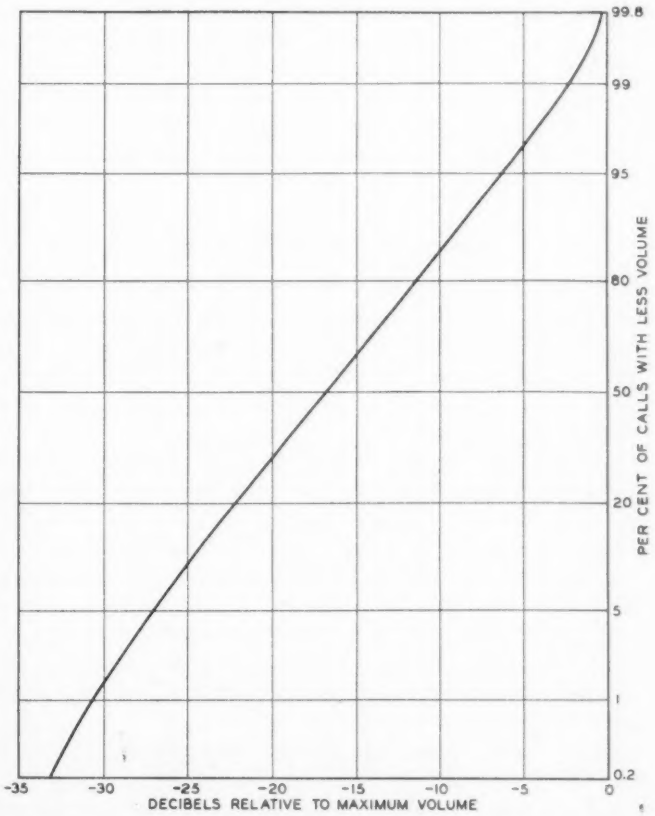


Fig. 1—Volumes of 950 local subscribers at New York transatlantic switchboard, January–April, 1931.

overloading appreciably impairs the signal quality or endangers the life of the equipment. It is an economic limit set by the cost of building equipment of greater load capacity. The lower limit of volume is set by the combination of the amount of attenuation and the amount of interference in the system such that the signal should not be appre-

ciably masked by noise. This also is ordinarily an economic problem depending on the cost of lowering the attenuation or of guarding against external interference. In some cases, however, this limitation is a physical one. A striking case is that of radio transmission in which we have no means of controlling the attenuation of the electromagnetic waves in transit to the receiving station. They may arrive at levels below those of thermal [3, 4] noise in the antenna and other receiving apparatus. Thus, even in the absence of static there is a definite useful lower limit to the received and hence the transmitted volume. In such cases the problems raised by the spread in signal intensities become a matter of particular importance. Radio telephony was therefore one of the fields of use particularly in view for the development of the device to be described.

Effect of Volume Control

Until recently the only method in use for reducing the range of signal intensities on radio circuits was a special operating method for constant volume transmission. At each terminal the technical operator, with the aid of a volume indicator, adjusted the speech volume going to the radio transmitter to that maximum value consistent with the transmitter load capacity.

Referring to Fig. 2, we have a diagram showing the normal relation of input to output intensities of a zero loss transducer as given by the diagonal line. Points $A_{\max.}$ and $A_{\min.}$ on this line indicate the extreme values of signal intensities for sustained loud vowels covering a volume range of 40 db. The effect of the volume adjustments made by the technical operator is to bring all the applied volumes to a single value indicated by point B in Fig. 2. The value of B could be any convenient intensity. Here it is set at a value determined by transmission conditions in the line between the technical operator's position and the radio transmitter.

As the technical operator has reduced the strongest volumes 5 db and increased the weakest volumes 35 db, the result of this volume control is to increase the volume range which the circuit can handle by 40 db. It is possible to make this adjustment for two-way transmission in the case of radio circuits without danger of singing because of the use of voice-controlled switching arrangements [5] which permit transmission in only one direction at a time. By this method of operation volumes initially strong or weak are delivered to the distant receiving point with equal margins relative to interference and the transmission capacity of the whole system is thereby improved.

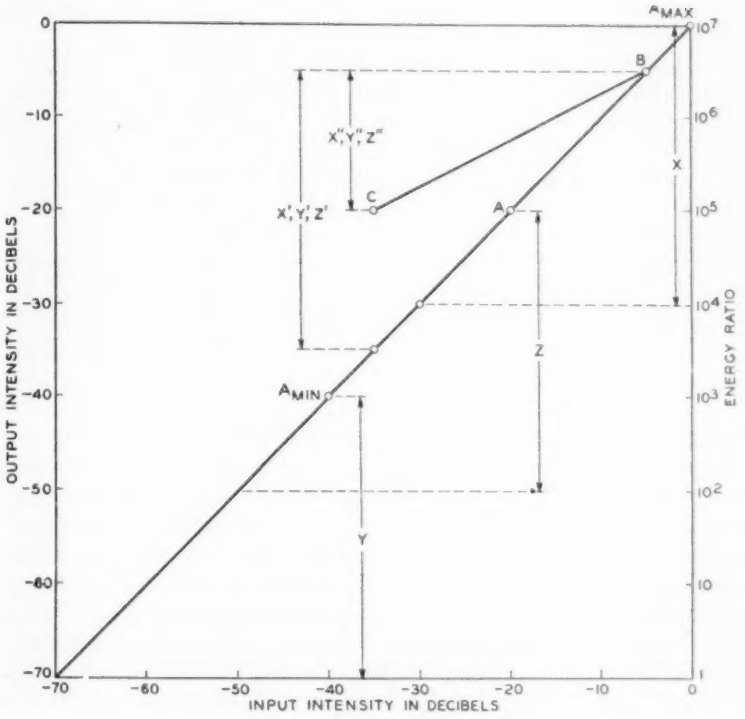


Fig. 2—Range control.

Intensity Range at Constant Volume

However, even with speech adjusted to constant volume at the transmitting point there are large variations in signal intensity from syllable to syllable and within each syllable. For example, the energy of some consonants as compared with the stronger vowels is down about 30 db. The importance of the weaker sounds is brought out by the fact that in the case of commercial telephone sets a steady noise 30 db below the energy in the strongest parts of the speech syllables produces an appreciable impairment in transmission efficiency. It is accordingly desirable to maintain transmission conditions such that generally more than this range is kept free from the masking effect of noise. This range of intensities within the syllable is also of importance in the operation of the voice-controlled switches used in the radio system. The sensitivity spread between a voice operated relay which

just operates on the crests of loud syllables and one which operates sufficiently well not to clip speech is also about 30 db.

Considering on Fig. 2 that the coordinates are in terms of the average r.m.s. value over a period of time small compared with the time of a syllable, there is a spread of at least 30 db in signal intensity extending down from the maximum for each talker. Thus for the weakest talker this spread is indicated by the bracket Y and for the strongest, by X. Any other talker, as Z, falls somewhere in between. After manual control of volume this spread of intensities is represented by the bracket X', Y', Z' for all talkers. This residual spread makes desirable a means for further compressing the range of intensities in the speech signals so that the weaker parts of sound are transmitted at a higher level without at the same time raising the peak values of speech and so overloading the transmitter.

Types of Compression Systems

This problem can be approached in several ways. One, for instance, is from the frequency distortion standpoint. As many of the weaker consonants have their chief energy contribution in the upper part of the speech band, a simple equalizer which relatively increased the energy of the higher frequency consonants before transmission and another which restored the frequency energy relations after transmission should be found of value. Tests have confirmed this expectation to some degree. Unfortunately, the best type of equalizer depends upon the type of subscriber station transmitter, so that in general only a compromise improvement can be obtained.

Another general method of approach is that of amplitude distortion in which the weaker portions of the syllable are automatically increased in intensity in some inverse proportion to their original strength. The manual control of volume described above may be considered the genesis of this method. Early suggestions [6] included the use of an auxiliary channel such as a telegraph channel for duplicating the control operations in the reverse sense at the receiving end, thus restoring the original energy distribution. Another early suggestion along this line was made by George Crisson of the American Telephone and Telegraph Company.[7] If a voltage be applied to a circuit consisting of a two-element vacuum tube (with a parabolic characteristic) in series with a large resistance, the instantaneous voltages across the tube are approximately the square root of corresponding voltages applied. Thus a voltage originally 1/100 of the peak voltage can be transmitted at an intensity of 1/10 of the peak or ten times its original intensity. If the instantaneous energy is expressed on the logarithmic

or db scale, the energy range is then cut in half. Such a device may be called an instantaneous compressor. At the distant end a circuit which is simply the inverse of that at the transmitting end is used. The output voltage is taken off of a low resistance in series with a parabolic element, thus restoring the signal substantially to its original form. This circuit may be called an instantaneous expandor. This scheme was successfully tested in the laboratory but unfortunately possesses a very serious limitation for practical application in the telephone plant. This is due to the fact that, to properly maintain the characteristics of the compressed signals, a transmission band width without appreciable amplitude or phase distortion of about twice the normal proved necessary.

The Compandor

The principle of the present device is the use of a rate of amplitude control for the compressing and expanding devices intermediate between manual and instantaneous control which may be considered approximately as a control varying as a function of the signal envelope.[8, 9] Such a modulation of the original signal in terms of itself does not appreciably widen the frequency band width of the modified signal as compared with the original signal. The transmitting device is called the compressor; the receiving device, the expandor; and the complete system, the compandor.

The functional behavior of a typical compressor may be considered with reference to the simplified schematic circuit No. 1 of Fig. 3.

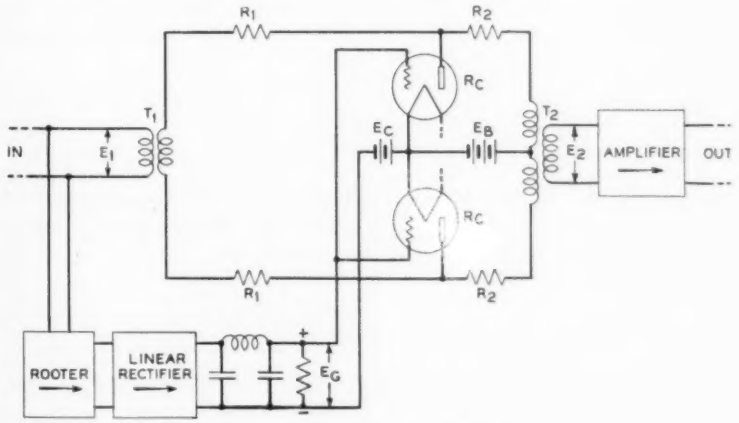


Fig. 3—Compressor circuit No. 1.

This circuit is of the forward-acting type; that is, the control energy is taken from the line ahead of the point of variable loss. The variable loss consists of a high impedance pad connected in the circuit through two high ratio transformers T_1 and T_2. The high resistances R_1 and R_2 are shunted by a pair of control tubes connected in push-pull. The push-pull arrangement is desirable for two reasons. It reduces the even order non-linear distortion effects caused by the shunt path on the transmitted speech and it balances out the control impulse and unfiltered rectified speech energy from the control path which might otherwise add distortion to the speech. The impedances of these tubes are controlled by the control voltage E_G, which is roughly proportional to the envelope of speech energy and which is derived from the line through a non-linear or "rooter" * circuit, a linear rectifier and a low-pass filter which may have a cutoff frequency in the range 20 cycles to 100 cycles. In the following analysis it is assumed that the delay due to this filtering is negligible:

Let E_1 = r.m.s. speech voltage at input

and E_2 = r.m.s. speech voltage at output in same impedance

R_C = a-c. impedance of control tubes.

Now if R_C is kept small compared to the pad impedance, we have approximately

$$E_2 = k_1 E_1 R_C. \tag{1}$$

Let E_G be the control voltage applied to the grids of the control tubes. With the plate voltage E_B just neutralized by the steady biasing grid voltage E_C, then only E_G may be considered as determining the space current and we may assume ideally that the space current

$$I_B = k_2 E_G{}^s.$$

Then

$$R_C = \frac{dE_B}{dI_B} = \mu \frac{dE_G}{dI_B} = \frac{1}{k_3 E_G{}^{s-1}}, \tag{2}$$

wheres s is determined by tube design and the ks are constants for constant μ tubes. For variable μ tubes equation (2) can be used to set requirements on the tube design.

From (1) and (2)

$$E_2 = \frac{k_1 E_1}{K_3 E_G{}^{s-1}}. \tag{3}$$

Now let the rooter be a non-linear circuit such that the instantaneous voltage is the tth root of E_1. After rectification and filtering we

* So called because the output is a root of the input; see equation (4).

shall have approximately

$$E_G = k_4 E_1^{1/t}. \tag{4}$$

From (4) and (3) we have

$$E_2 = \frac{KE_1}{E_1^{(s-1)/t}} = KE_1^{(t-s+1)/t}. \tag{5}$$

If $t = s = n$

$$E_2 = KE_1^{1/n}. \tag{6}$$

Now if the input voltage be increased by a factor x, the input increment in db will be 20 log x. The new output will be

$$E_2' = K(xE_1)^{1/n}.$$

The increment in output in db will be

$$20 \log \frac{E_2'}{E_2} = 20 \log x^{1/n} = \frac{20}{n} \log x.$$

The ratio of the output increment to the input increment in db is $1/n$ and the device is said to have a compression ratio of $1/n$. In other words, the per cent change in relative speech voltages in passing through the compressor is the same at all points in the intensity range. In the general form of this circuit, t and s need not be equal to secure a particular value of $1/n$.

In Fig. 4, Compressor Circuit No. 2 is shown, a backward-acting type of circuit. In this circuit the control tubes can be used to per-

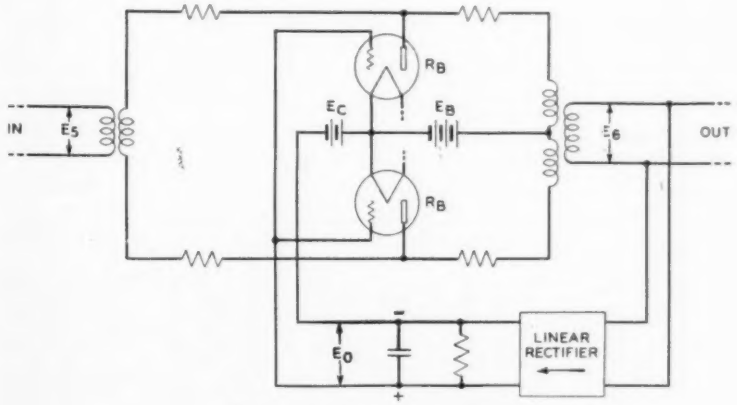


Fig. 4—Compressor circuit No. 2.

form the function of the rooter in circuit No. 1 when $s = t = n$. We may write for this circuit

$$E_6 = k_1 E_5 R_B,$$
$$E_0 = k_2 E_6,$$
$$R_B = \frac{1}{k_3 E_0^{n-1}} = \frac{1}{k_4 E_6^{n-1}},$$
$$E_6 = \frac{k_1 E_5}{k_4 E_6^{n-1}} = K E_5^{1/n}, \tag{7}$$

which is the same as equation (6) for circuit No. 1.

In Fig. 5 is shown the Expandor Circuit. If the resistances r are

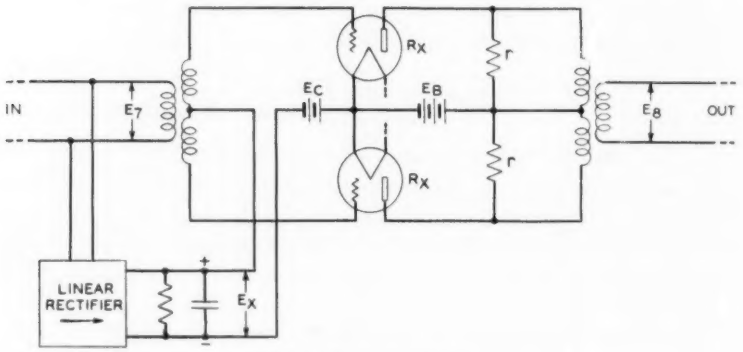


Fig. 5—Expandor circuit.

kept small compared with those of the control tubes, we may write

$$E_8 = \frac{k_1 E_7}{R_x},$$
$$E_x = k_2 E_7,$$
$$R_x = \frac{1}{k_3 E_x^{n-1}} = \frac{1}{k_4 E_7^{n-1}},$$
$$E_8 = K E_7 E_7^{n-1} = K E_7^n. \tag{8}$$

This relation is just the inverse of that given in equations (6) and (7). The increment ratio in db of output to input is n and the expansion ratio may be said to be n. When a compressor and expandor having the same value of n in their indices are put in tandem, the final output and input intensity ranges are the same. However, between the compressor and expandor the range of signal intensities, whose

rate of change is not faster than the usual syllabic envelope, is $1/n$ in terms of db. In terms of voltage ratios the intermediate signal intensities are proportional to the square root of their original values if n equals 2, the cube root if n equals 3, etc.

The ideal relations postulated above cannot all be met in the physical design of the circuits. The indices s and t must be the dynamic characteristics of the tube and circuit and can be held to constant value only over limited ranges of operation. Equation 2 is only approximately true as some space current is permitted to flow when no speech is passing; otherwise, impractical values of control impedances would be involved. However, they do serve to illustrate the functional operation and can be approximated sufficiently well in commercial equipment for useful amounts of compression and expansion. Figure 6 shows experimental steady-state input versus output characteristics for devices built to have a compression ratio of 1/2 and an expansion ratio of 2.

The compressor is seen to operate substantially linearly over a 45 db range of inputs and the expandor over a 22.5 db range. This is about as much range as can be secured conveniently from a single stage of vacuum tubes. As such ranges would be entirely insufficient to handle the seventy odd db range at speech intensities, it is necessary to control volumes to a given point before sending through these devices, rather than compress or expand first and then control. The range is adequate, however, to take care of the range of signal intensities for commercial speech at constant volume.

Effect of Compandor

The compressor curve of Fig. 6 indicates that, when the input is 15 db above 1 milliwatt, the compressor gives no gain or loss. If the levels are adjusted so that this point corresponds to the intensity at point B on Fig. 2, then the line BC indicates the controlled intensities corresponding to the assumed 30 db spread of speech controlled to constant volume. The new range of intensities as indicated by the bracket X'' Y'' Z'' is now finally reduced to about 15 db. Tests show that a volume indicator on the output of the compressor reads from 1 to 2 db higher than on uncompressed speech at its input. Compressed speech sounds slightly unnatural but the effects of compression upon articulation in the absence of noise are negligible.

In considering the action of the expandor it is important to note that all of the improvement in signal-to-noise ratio is put in by the compressor. Considering any narrow interval of speech the insertion of the expandor does not change the signal-to-noise ratio. The de-

sirability of using it depends on other reasons. First, it restores the naturalness of the speech sounds. Second, the apparent magnitude of the noise is greatly reduced since noise comes in at full strength only when speech is loudest and is reduced by the loss introduced by the expandor at times when the energy is low between syllables. When no speech is being transmitted, noises up to a certain limit, which

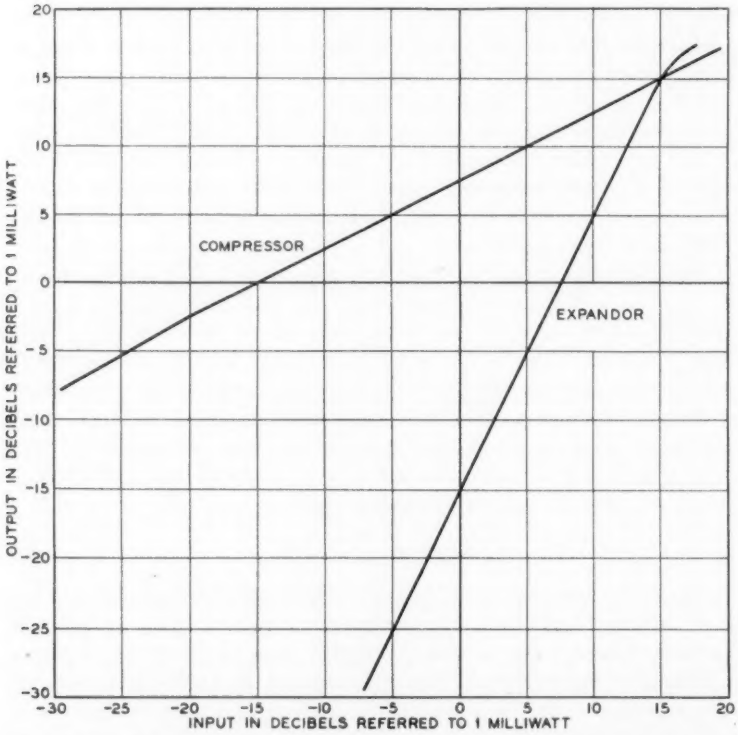


Fig. 6—Experimental input *vs.* output characteristics (1000 cycles steady state).

corresponds to the maximum energy in received speech, are reduced in varying amounts from about 20 db to zero depending on their value.

When speech is present the effect of the expandor is determined by the sum of the instantaneous speech and noise voltages, so that the effect on the noise, whether it is large or small, is determined largely by the existing speech intensity. For a circuit having somewhere near the limit of static, the use of the compandor allows on the average 5 db more noise than when it is not used. When the noise is less than

this limit, somewhat greater improvements are obtained from the compandor, ranging up to at least 10 db.

The particular values of compression and expansion ratio were chosen initially for the relative ease in the design of the system with commercially available vacuum tubes whose characteristics closely approximated a parabola. Tests of the equipment have shown that this degree is sufficient for present telephone circuit intensity range requirements. Increasing the amount of compression is limited by increase in quality distortion and by increased variation in the intensity of radio noise as heard by the listener. A noise which is constant at the input to the expandor varies on the output as the speech intensity changes. Also variations in attenuation equivalent between the compandor terminals are multiplied by the expandor. Herein lies a reason for having a constant compression and expansion ratio over the working range. If it were different at different intensities, attenuation changes would distort the reproduced speech as well as appearing as a somewhat increased change in intensity. This change in intensity is n times the attenuation change in front of the expandor in db.

The degree of compression may obviously be controlled in a variety of ways: such as, using different values for the indices s and t, applying control voltages upon more than one variable stage in tandem, the use of variable μ vacuum tubes, etc. The circuits as shown use variable shunt control for the compressor and variable series control for the expandor. Either or both may be changed to the other by inverting the polarity of the control potential and properly designing the rectifier characteristics of the control circuits.

There are two major sources of possible speech distortion which must be considered in the design and use of these devices in addition to those ordinarily present. The first is due to the non-linear characteristics of the vacuum tubes used for controlling. The even order distortion terms are largely balanced out by using two tubes in a push-pull arrangement. The remaining distortion is minimized by having speech pass through the control tubes at a sufficiently low level. In the operating ranges for the device shown on Fig. 6, the harmonics of a single-frequency tone are 30 db or more below the fundamental.

The second major source of distortion is the time lag in the control circuits due to the presence of the filters after the linear rectifier. However, with a complete compandor circuit using the compressor circuit No. 1, it was found on careful laboratory tests with expert listeners that it was almost impossible to distinguish whether the device

was in or out of circuit. Furthermore, distortion of this type is largely eliminated when compressor circuit No. 2 is used. In that case it will be noted that, if the two terminals are connected by a substantially distortionless transmission system, the identical control circuits of the two devices receive identical operating voltages. As the gain changes put in are reciprocal and occur now with equal time lag, the deviations from ideal compression are virtually counterbalanced by the inverse deviations from ideal expandor action. In Fig. 7 are shown

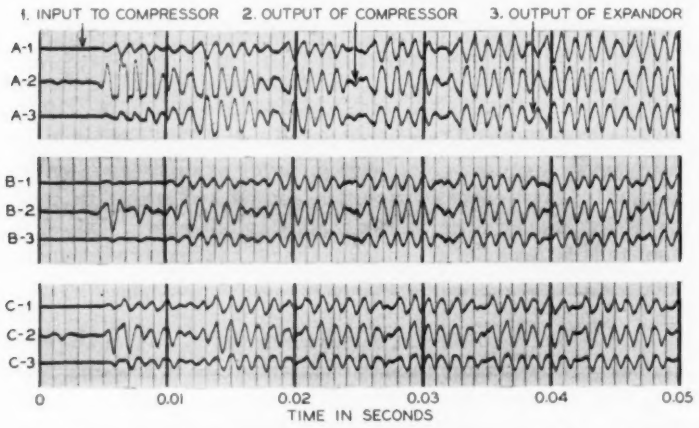


Fig. 7—Operation of compandor on beginning of word "bark." *A*. Compressor circuit No. 1. *B*. Compressor circuit No. 2 with low-pass filter in control circuit. *C*. Compressor circuit No. 2 without filter.

oscillograms taken of the first part of the word "bark." Each record shows the intensity changes before the compressor, between the compressor and expandor and on the output of the expandor.

Application to Transatlantic Circuit

A compandor system has been in service on the New York–London long-wave radiotelephone circuit since about July 1, 1932. At first compressor circuit No. 1 was used, and later a change was made to compressor circuit No. 2. Figure 8 is a photograph of the experimental installation at New York. It occupies about five feet of standard relay rack space. The blank panel shown in the photograph indicates the saving of apparatus resulting from the change to compressor circuit No. 2. Figure 9 is a schematic diagram showing the method of inserting the compressor and expandor in the radio telephone terminals at each end of the circuit. Since the two ends are similar, only one

end is shown. The compandor circuits are indicated in their relation to the subscriber, the toll switchboard, the vodas and privacy apparatus, and the radio transmitter and receiver.[5] A meter located at the point designated *A* would indicate the full range of applied volumes, at *B*, the controlled volumes and at *C*, the compressed speech signals.

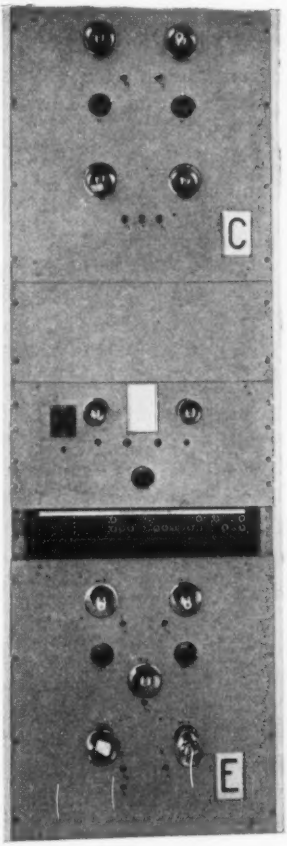


Fig. 8—Experimental installation of compandor at New York.

When the United States subscriber talks, electrical waves set up by his voice pass over a wire line to the toll switchboard. They then divide in a hybrid set; part of the energy is dissipated in the output of a receiving repeater and part is amplified by a transmitting repeater whose gain is controlled by noting the reading of a volume indicator at

B and adjusting a potentiometer ahead of the transmitting repeater. The waves then act on the vodas which consists of amplifier-detector, delay circuit and relays for switching the transmission paths in such a manner as to prevent echoes, singing and other effects. When in the transmitting condition, the vodas is arranged to have zero loss so that the waves impressed on the compressor are practically the same as at *B*. The waves put out from the compressor are then sent through the privacy apparatus, the output of which is then sent over a wire line to the radio transmitter. The radiated waves are picked up by the distant radio receiver, amplified and transformed into voice-frequency energy which passes over a wire line to the terminal at the distant end.

The path of received waves in either terminal may be traced in the lower branch of the circuit shown on Fig. 9. After being made intel-

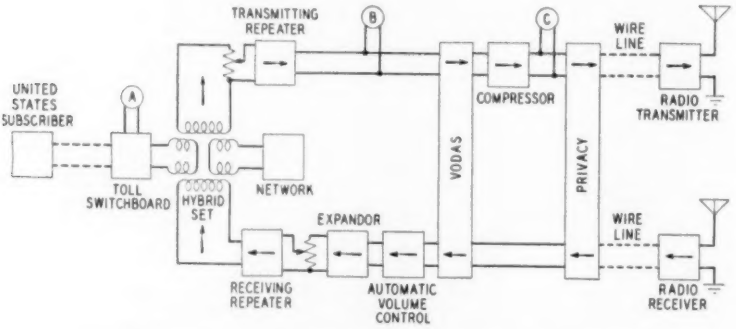


Fig. 9—Compandor applied to one end of a radio telephone circuit.

ligible by passing through the receiving privacy device, the compressed incoming waves are sent through the vodas into an automatic volume control and then into the expandor. The expanded waves are sent through a receiving repeater from whose output the amplified waves pass into the hybrid set, part being dissipated in the network and the other part going through the toll switchboard to the subscriber. Due to imperfect balance between the subscriber's line and the network, a portion of the received energy is transmitted across the hybrid set and amplified by the transmitting repeater. This echo might operate the transmitting vodas under certain conditions. For this reason a potentiometer is inserted in the receiving branch of the circuit so as to reduce the echo, and consequently the received volume, so that false operation of the transmitting vodas is prevented.

Results of Compandor Operation

The effectiveness of the compandor in service depends not only on its ability to reduce noise but also on its relation to the other characteristics of the circuit. Tests in the laboratory and on the long-wave transatlantic circuit have indicated that the presence of the compandor does not affect the quality appreciably, provided compressor circuit No. 2 is employed and provided the compression in the circuit itself is not serious. Delay distortion can be tolerated up to about the same amount as when no compandor is used. Frequency changing for privacy purposes is not materially affected by the compandor.

The expandor increases the transmission variations in the circuit exactly as it increases the voltage range of the waves applied to it. It is therefore necessary to guard against excessive variations in the overall circuit including the wire line extensions as well as the radio links. At the New York terminal there has been installed an automatic volume control operated from received speech signals which performs this function.

The received volume is limited by incoming waves which do not operate the receiving side of the vodas but which return as echoes from the land line to cause false operation of the transmitting side. The compressor increases these weak waves so that they are better able to operate the receiving side of the vodas, and the expandor effectively increases the stronger waves relative to the weak. This results in more received volume being delivered to the two-wire terminal than when the compandor is not used. The overall improvement in volume delivered to the subscriber varies with the noise, being greatest when the noise is low.

Summary

The allowable increase of about 5 db in noise before reaching the commercial limit increases the time when the circuit can be used for service. The increased circuit time is greatest in the seasons of the year when it is needed the most.

For conditions of moderate disturbances now classed as commercial, a reduction of the noise transmission impairment to very low values is accomplished by the compandor.

The improvement in the vodas operation results in delivering substantially higher volumes to the subscribers.

The beneficial effect of the compandor might alternately be applied to a reduction of transmitter power.

References

1. "The New York–London Telephone Circuit," S. B. Wright and H. C. Silent, *Bell System Technical Journal*, Vol. VI, pp. 736–749, October, 1927.
2. "Speech Power and Its Measurement," L. J. Sivian, *Bell System Technical Journal*, Vol. VIII, pp. 646–661, October, 1929.
3. "Thermal Agitation of Electricity in Conductors," J. B. Johnson, *Physical Review*, Vol. 32, pp. 97–109, July, 1928.
4. "Thermal Agitation of Electric Charge in Conductors," H. Nyquist; presented before the *American Physical Society*, February, 1927, and published in *Physical Review*, Vol. 32, pp. 110–113, July, 1928.
5. "Two-Way Radio Telephone Circuits," S. B. Wright and D. Mitchell, *Bell System Technical Journal*, Vol. XI, pp. 368–382, July, 1932, and *Proceedings of The Institute of Radio Engineers*, Vol. 20, pp. 1117–1130, July, 1932.
6. U. S. Patent 1,565,548, December 15, 1925, issued to A. B. Clark.
7. U. S. Patent 1,737,830, December 3, 1929, issued to George Crisson.
8. U. S. Patent 1,738,000, December 3, 1929, issued to E. I. Green.
9. U. S. Patent 1,757,729, May 6, 1930, issued to R. C. Mathes.

The Effect of Background Noise in Shared Channel Broadcasting

By C. B. AIKEN

The interference which occurs in shared channel broadcasting consists of several components of different types. Of these the program interference is usually the most important in the absence of a noise background, while if a strong noise background is present another component, which may be called flutter interference, predominates.

A simple theory of the flutter effect is developed and it is shown that its importance is dependent upon the type of detector employed. If manual gain control is used, flutter may be greatly reduced by the use of a linear rectifier. However, if automatic gain control is used this superiority of the linear detector cannot be realized and flutter is bound to be troublesome.

The results of experimental studies of the various types of interference are given and a comparison is made of the relative importance of flutter and program interference. The effects of the type of detector used and of the width of the received frequency band are observed. It is evident from these studies that improvements in the size of the lower grade service areas of shared channel stations might be obtained by close synchronization of the carrier frequencies, even though different programs are transmitted.

THE regulation requiring that carrier frequencies be maintained to within fifty cycles of their assigned values has resulted in the practical disappearance from shared broadcast channels of the heterodyne whistle, that most pernicious of all types of radio interference. Consequently, it is now unnecessary to have so large a ratio of the field strength of the desired signal to that of the undesired as was the case before the banishment of the high pitched squeal. Nevertheless, the field strength ratio which is necessary to permit of satisfactory reception on shared channels is still much higher than we should like it to be, and interference still abounds.

A very common type of interference is that which manifests itself as a fluttering or heaving sound, often very unpleasant in character. This phenomenon is caused by the periodic rise and fall of the background noise (static, R. F. tube and circuit noise, etc.) as the weak interfering carrier wave swings alternately in and out of phase with the carrier from the stronger station. In the complete absence of a noise background, program interference, or "displaced sideband interference" [1] as it may be called, is more troublesome than are flutter effects. Consequently, it is in regions other than the high grade service areas of shared channel stations that flutter effects are most annoying. In such regions they occur most prominently when the

[1] "The Detection of Two Modulated Waves Which Differ Slightly in Carrier Frequency," *Proc. I. R. E.*, January, 1931, and *Bell. Sys. Tech Jour.*, January, 1931.

frequency difference of the desired and interfering carriers is only a few cycles per second. As this difference is increased the flutter is transformed into a more sustained sound, rather harsh in character, and as it is still further increased a low growl appears which becomes more objectionable as it rises in frequency. The pitch of this growl cannot exceed 100 cycles unless one or both stations are violating the 50 cycle regulation. With the increasing use of very precise frequency control, heterodyne frequencies of a few cycles have become very common, and so, therefore, have flutter effects.

It has been pointed out in an earlier paper [2] that the magnitude of the flutter effect will depend upon the type of rectifier employed in the receiving set, and that it will be very much more objectionable when a square law detector is used than when a linear detector is employed. This is to be expected, since in the former case the audio-frequency output of the receiver will be proportional to the amplitude of the incoming carrier, while in the latter case the output will be essentially independent of the carrier amplitude, provided over-modulation does not occur. However, these statements refer to the case in which automatic gain control is not used. When the receiver is equipped with automatic control, as in most better grade modern receivers, the superiority of the linear detector is nullified and a serious flutter may occur.

In addition to displaced sideband interference and flutter, trouble may arise from distortion of the desired program by the action of the interfering carrier. One or both of the first two types of interference are likely to occur at lower field strength ratios than is the last, but at higher levels of the undesired carrier all three types are of importance and combine to degrade the quality of reception. In this paper, studies of all these types will be reported. Audible beat interference will not be discussed since it has been considered in other papers and, as just mentioned, is much less important than it used to be.

Theoretical Estimation of Flutter Effects

As has already been stated, the flutter effect is due to the rise and fall of the level of the noise background with variation in the effective amplitude of the impressed carrier. In order to study this effect, let us suppose that there are impressed upon the detector a component of radio frequency noise which may be represented by $N \cos (\omega + n)t$, and a desired carrier $E \cos \omega t$. $n/2\pi$ is assumed to be an audio-frequency.

If a square law, or quadratic, detector is employed, the audio-

[2] "Theory of the Detection of Two Modulated Waves by a Linear Rectifier." *Proc. I. R. E.*, Vol. 21, pp. 601–629, April, 1933.

frequency output will be proportional to the audio-frequency component of

$$[E \cos \omega t + N \cos (\omega + n)t]^2,$$

which is

$$EN \cos nt. \tag{1}$$

Now suppose that there is impressed, in addition to the desired carrier and noise component, a weak carrier $e \cos (\omega + u)t$. The sum of the strong and weak carriers may be conveniently regarded as a single wave of amplitude

$$(E + e \cos ut).$$

This may be substituted for the amplitude E in (1), giving for the noise output

$$EN(1 + K \cos ut) \cos nt \tag{2}$$

in which

$$K = e/E. \tag{3}$$

The noise which is heard will consist of a steady portion, the amplitude of which is proportional to EN, and another portion of variable amplitude which is proportional to $ENK \cos ut$.

The factors that determine the importance of the flutter are many and complex, but it seems likely that the most important of them is the ratio of the variable component of the noise output to the steady component. As long as the noise is loud enough to be obvious, this ratio should be a fairly good measure of the perceptibility of the flutter, and we shall venture to regard it as such. The experimental data to be reported later will bear out this assumption.

From (2) it is evident that the ratio mentioned is merely K, the ratio of the amplitude of the interfering carrier to that of the desired carrier. We shall call this ratio the "flutter factor" for the quadratic detector and designate it by F_Q.

$$F_Q = K = e/E. \tag{4}$$

It is interesting that F_Q is independent of the amplitude N of the high frequency noise.

It is possible to derive a similar factor, giving the ratio of the variable to the steady components of noise, for the linear detector. From equations (70*a*) and (71) of the paper [2] already mentioned it follows that the flutter factor for the linear detector, at low modulations of the desired wave, is

$$F_L = \frac{Ne}{4E^2} = \frac{kK}{4}, \tag{5}$$

in which $k = N/E$.

F_L is seen to be dependent upon the strength of the high frequency noise as well as upon that of the interfering carrier. It is also to be noted that the flutter will be more serious with the quadratic than with the linear detector by a factor $4/k = 4E/N$, which is usually large.

This derivation of F_Q and F_L on the basis of a single frequency noise component serves to indicate important differences between the two types of detector and to show how the flutter changes with the noise level and with the ratio of the incoming carrier amplitudes. In any practical case the noise field would consist of numerous frequency components, but it is reasonable to expect that the proportionalities expressed in (4) and (5) would still hold. However, the absolute values of N and K at which the flutter becomes detectable must be determined experimentally and may be expected to depend upon the width of the received frequency band.

In the foregoing derivations it has been assumed that there is no automatic volune control in the receiving set. A brief examination of the effect of such a device will now be made.

Action of an Automatic Volume Control

The comparative freedom from flutter effects which has been noted in the case of the linear detector may be regarded as due to the fact that the audio-frequency output of such a detector is independent of carrier amplitude over a wide range. If automatic volume control is used in the receiving set, the amplitude of the carrier wave will be maintained practically constant at the input terminals of the detector. If the effective carrier amplitude impressed upon the antenna undergoes a periodic fluctuation, due to very low frequency heterodyning between the two stations, the gain of the radiofrequency amplifier will undergo cyclic variations, so as to keep the carrier constant at the detector. Obviously this will cause a fluctuation in the amplitude of the side-bands, be they due to noise or program.

From this it is evident that, on the one hand, flutter effects in the presence of a noise background will usually be of minor importance if a good linear rectifier is employed in conjunction with a manual volume control; while, on the other hand, these effects may become extremely objectionable if automatic volume control is used. Because of the prevalent use of AVC in modern radio receivers the low flutter characteristics of the linear detector cannot be generally employed to reduce flutter interference on shared channels.

In the case of the square law detector, the output is proportional to the product of the amplitudes of the carrier and side frequencies. At first glance it might seem that the use of automatic volume control

should reduce the flutter effects, since it would iron out the variations in carrier amplitude impressed upon the detector. However, it is evident that this stabilization of the carrier will be exactly offset by the variation imposed upon the sideband amplitudes, and that consequently the flutter effects should be as evident when a normally functioning automatic volume control is used as they are in the case of manual control.

A perfectly functioning automatic volume control should make flutter effects approximately independent of the type of detector employed when the beat frequency is of the order of 2 or 3 cycles. However, at some of the higher frequencies, of the order of 20 to 40 cycles, the control will function with reduced efficiency, and at still higher frequencies will not function at all. Consequently, in this intermediate range the gain control may have some special effect and may make the flutter either worse or better than it would be with the same type of detector and manual control.

Experimental Studies

Equipment Used in the Study of the Effects of a Noise Background

A laboratory investigation was made of the interference between two waves of slightly different carrier frequency. A block schematic of the equipment used is shown in Fig. 1.

A modulated signal could be received from Station WABC, or, by throwing the switch *S*, it was possible to obtain an unmodulated carrier from a Western Electric No. 700A Oscillator, which is of very great frequency stability.[3] Whichever signal was used was fed through an impedance matching transformer to a radio frequency attenuator. The output of this attenuator was fed into the grid of one tube of a mixing amplifier. As indicated in the drawing, this amplifier consists merely of two shield grid tubes having a broadly tuned common plate circuit load.

The other tube of the mixing amplifier was energized, through a second radio frequency attenuator, by an unmodulated carrier derived from a crystal controlled laboratory oscillator of the same type as that which served as an alternative to WABC. This oscillator was part of a Western Electric No. 1A Frequency Monitoring Unit.[4] The monitor includes arrangements for measuring frequency differences between the oscillator included within it and an external source. In this case the external source was WABC, or the alternative carrier. The energy

[3] O. M. Hovgaard, "A New Oscillator for Broadcast Frequencies," *Bell Laboratories Record*, 10, 106–110, December, 1931.

[4] R. E. Coram, "A Frequency Monitoring Unit for Broadcast Stations," *Bell Laboratories Record*, 11, 113–116, December, 1932.

required by the frequency measuring device was supplied through a tuned buffer amplifier.

The voltage developed across the tuned circuit of the mixing amplifier was measured by a conventional form of vacuum tube voltmeter. By setting one attenuator at a very high loss, the magnitude of the signal supplied through the other could be measured, and the process then reversed. If the two signals were adjusted so as to give equal amplitudes across the tuned load, then any desired carrier ratio could be obtained by adding a known loss in one attenuator.

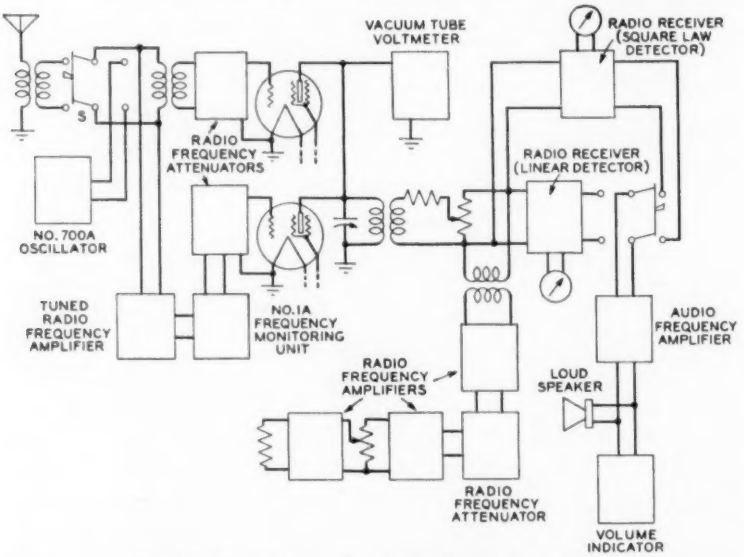


Fig. 1—Schematic circuits of experimental setup.

The mixing amplifier fed a shielded transmission line which included an adjustable pad. The line supplied energy to either of two radio receivers, one of which contained a square law and the other a linear detector. The output of the receiver was monitored on a loud speaker and also on a volume indicator. Meters were provided for indicating the change in direct current flow in the detector circuit of both receivers.

In order to study the effects of a noise background, a noise source of constant and controllable level was required. Furthermore, it was desirable that the noise be of a type frequently encountered in practice. The thermal noise generated in a high gain amplifier seemed to be suitable. Consequently, there were connected in cascade two amplifiers having a gain of approximately 44 db each, over the entire broad-

cast band. The output of the second of the units was fed through a radio frequency attenuator to the grid of a single stage amplifier, the output circuit of which contained a step-down transformer bridged across the transmission line feeding the radio receivers. With zero loss in the attenuator the noise energy fed to the line was ample for the purposes of the present study.

An additional description of some of the pieces of equipment used in the foregoing set-up may be of interest.

Source of Constant Unmodulated Carrier Frequency

The oscillator contained in the No. 1A Frequency Monitoring Unit is of unusual frequency stability. The piezo-electric crystal is mounted in a specially designed thermal insulating chamber which reduces the temperature fluctuations to an extremely small fraction of a degree. Voltage regulating equipment is included in the unit, giving further assistance in stabilizing the frequency. Detailed descriptions of the oscillator [3] and of the frequency monitor [4] have been published.

A similar oscillator is used as a control unit at Station WABC. Hence, it was expected that a very constant beat frequency could be obtained between that station and the local oscillator. The frequency of the latter was adjustable over a narrow range by means of a vernier condenser in the crystal circuit. Figure 2 shows a number of plots of

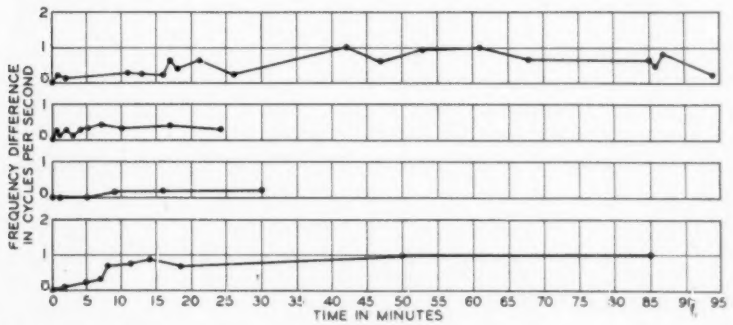


Fig. 2—Beat frequency between WABC and Western Electric No. 1A Frequency Monitoring Unit.

the beat frequency against time. These curves indicate an extremely slow drift, and experience has shown that the beat frequency would hold to within 0.4 cycle over a period of at least five minutes, and usually considerably longer. This high stability greatly facilitated work which required a very small difference in frequency of the two carriers.

Radio Receivers

Both receivers were high fidelity (7000 cycles) units of the tuned radio frequency type. One of these was modified so that either manual or automatic volume control could be used, and the level impressed upon the detector was reduced so that it would function as a strictly square law device. The cathode resistor which normally furnishes a grid bias for the detector tube was replaced by a battery. This was necessary in order to prevent straightening out of the characteristic by degeneration at very low frequencies. In Fig. 3 is shown a plot of

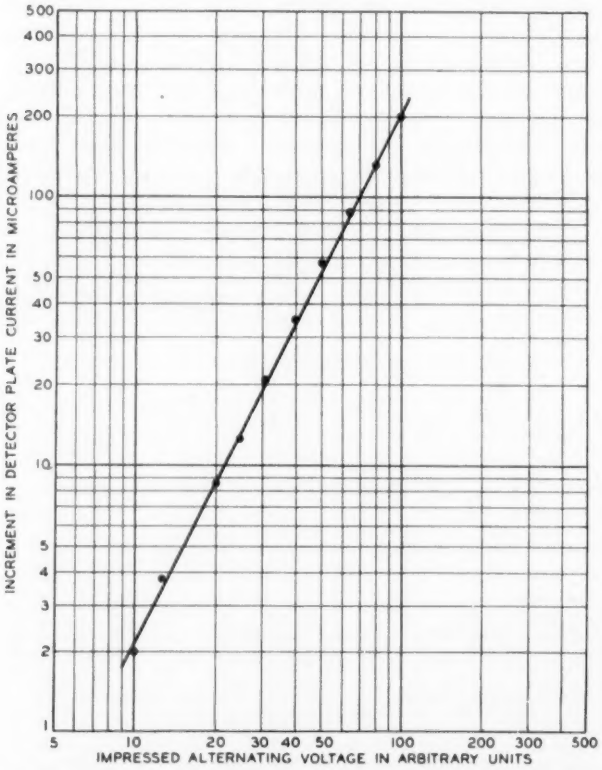


Fig. 3—Characteristic of radio frequency amplifier and square law detector.

the change in detector space current as a function of the impressed voltage. It will be observed that for increments of less than 200 μA the characteristic has a slope of two to one. All observations were

made at signal levels which were low enough to stay well within this range.

The other set was provided with a diode rectifier which functioned as a linear detector. In order to improve the linearity of the characteristic, an initial bias was used and was adjusted to obtain the best characteristic as indicated by the following test:

If a large unmodulated carrier is impressed on a linear rectifier, together with a much smaller unmodulated carrier, the beat frequency output should be independent of the amplitude of the larger carrier over a wide range. This phenomenon was observed experimentally and the initial bias was altered until the range, over which the large carrier could be adjusted without changing the output, was a maximum. In Fig. 4, the horizontal curve shows the magnitude of the

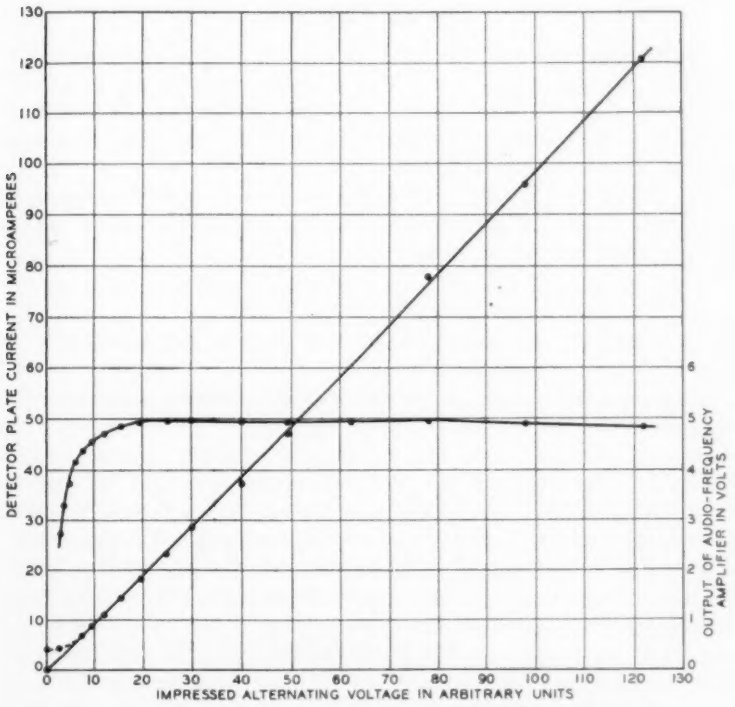


Fig. 4—Characteristics of radio frequency amplifier and linear detector.

audio-frequency output, while the sloping curve shows the direct current flowing in the detector circuit. The dashed curve is due to the

presence of the weak signal, while the solid one represents the effect of the large signal alone. The curves of this figure were taken with a bias of + 0.5 volt, which was found not to be critical.

The results of the experimental observations made with this detector were entirely in accord with theory, as will be discussed later, while similar observations made with a zero bias gave results which differed considerably from those predicted by the theory of the linear rectifier. Lack of the small bias caused a considerable departure from linearity, as was plainly evidenced by the fact that when it was absent the audio-frequency output due to the two carriers was by no means independent of the magnitude of the larger.

The tuned circuit in the mixing amplifier was so broad as to have an entirely negligible effect on the fidelity of the radio receivers.

Listening Conditions

In studying the effects of noise background some observations were made in the open laboratory, and a greater number in a partially deadened room 10 feet x 10 feet x 10 feet. The sound-proofing of this room was sufficient to keep out street noises and other extraneous disturbances of moderate intensity.

In determining the dependence of a given effect upon the magnitude of the carrier ratio, there was recorded that value of the ratio at which the effect was just perceptible.

Results of Experimental Work

A number of observations have been made with the intention of obtaining practical data on the characteristics of reception in the presence of a noise background, and with the purpose of checking the theoretical predictions already given. It has been pointed out that the flutter effects depend upon the type of detector which is employed and upon the ratio of the two carriers. If a square law detector is used the effect should be very nearly independent of the magnitude of the noise level, so long as it is within reasonable limits and does not either overload any of the equipment (including the ear of the listener) or fall so low as to be hardly noticeable. On the other hand, if a linear detector is employed, flutter effects should increase with the noise level. In either case the modulations of the two stations play no important part in determining the flutter effects except in so far as high modulations may temporarily mask them.

As a result of these considerations it was decided to employ unmodulated carriers for the greater part of the work. In order that a suitable level might be chosen, the strong carrier was first adjusted to

give the proper change in detector current. It was then modulated 30 per cent with a pure tone, and the gain of the audio-frequency output amplifier was adjusted until a fairly loud, but entirely comfortable, level was delivered to an observer placed about six feet in front of the loud speaker. The output level of the audio-frequency amplifier was read on a meter so that its gain might be checked later on.

The Linear Rectifier

The detector of a radio receiver was adjusted to have a linear rectifier characteristic in the manner just described and manual gain control was employed. In the first set of runs the carrier ratio was determined at which the flutter effect at low frequencies, or the carrier beat-note at higher frequencies, became just noticeable, the frequency being the variable. In Fig. 5 is shown a curve representing a number of

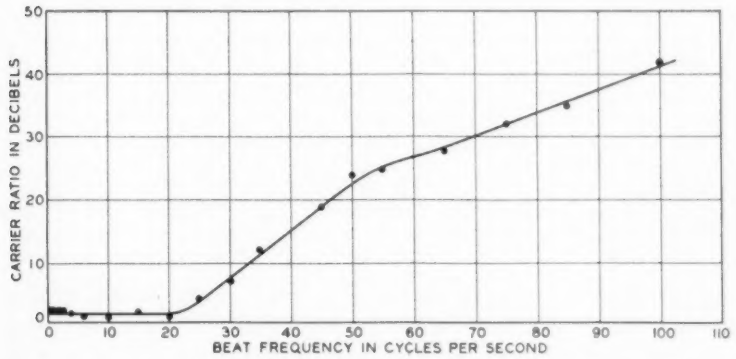


Fig. 5—Carrier ratio for perceptible flutter with a linear detector. Noise equivalent to 9.5 per cent modulation.

observations of this type. The noise level was constant at 10 db down from a 30 per cent modulated signal. By this it is meant that when the noise was impressed upon the receiver, together with a carrier the level of which had been fixed as described above, the audio-frequency output, as measured on a copper oxide level indicator, was 10 db below the audio output resulting from a 30 per cent modulation of the same carrier in the absence of noise.

A very interesting fact to be noted from this curve is that, for beat frequencies of less than about 20 cycles, the carriers must be very nearly equal before any flutter effect whatever may be detected. The average curve has been drawn through a value of 1.5 db. The observed values vary from this figure by not more than $\pm$ 0.5 db.

The right-hand portion of the curve is determined by the audibility of the beat-note, and its position will of course depend upon the masking effect of the noise background. Theory has indicated that the flutter frequency portion of the curve should drop with the noise level, but it is evident that, with such a small difference in carrier amplitudes as is indicated in the figure, the results would not be appreciably different were the noise level to be reduced. On the other hand a noise level which is down only 10 db from a 30 per cent modulated signal is equivalent to a modulation of nearly 10 per cent. This is an extremely objectionable noise level, so objectionable, in fact, that under the conditions of the tests it was very unpleasant to listen to. Consequently, it did not seem worth while to run curves similar to that of Fig. 5 for a number of different noise levels. Instead, a set of observations was made with a fixed carrier frequency difference of 2 cycles and a variable noise level. The results are indicated by the lower curve in Fig. 6.

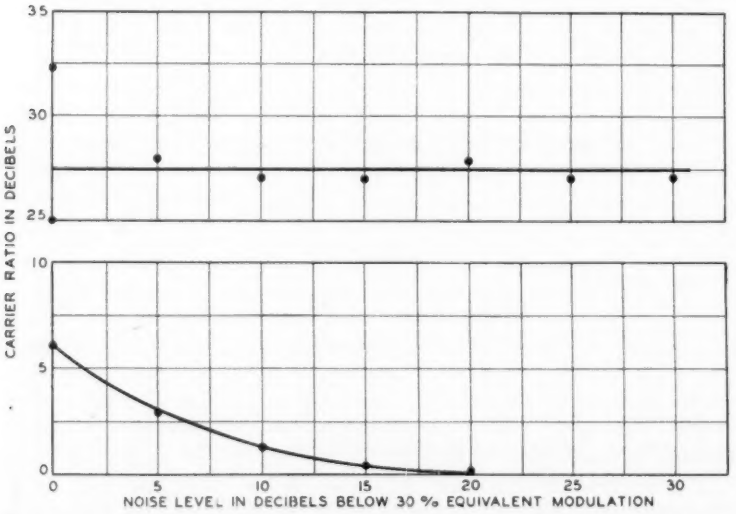


Fig. 6—Carrier ratio for perceptible flutter as a function of noise level. The upper curve is for the *square law* and the lower curve for the *linear detector*.

With a noise level equivalent to a 30 per cent modulation, a carrier ratio of only 2 : 1 is necessary to reduce the flutter to a barely detectable amount. At low noise levels, down 20 db or more from 30 per cent, the flutter could hardly be detected but there was noticeable a "bumping" sound which was due to the rather violent motion of the cone of the loud speaker at a frequency of 2 cycles. This was partially eliminated

by inserting a capacity in series with the voice frequency circuit of the speaker, but even when greatly reduced the bumping was detectable and was more important than any flutter which may have been present.

The Square Law Rectifier

Observations similar to those just discussed were made with a square law detector. In Fig. 7, the ordinates represent the carrier ratio neces-

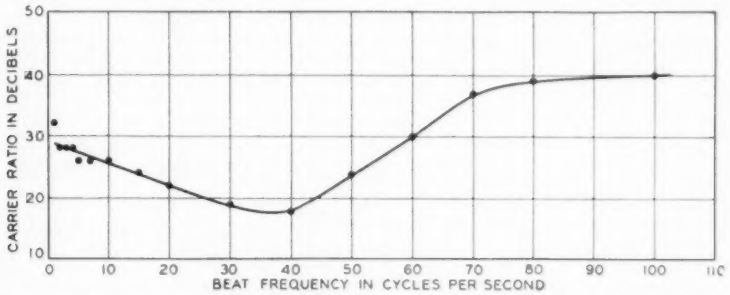


Fig. 7—Carrier ratio for perceptible flutter with a square law detector. Noise equivalent to 9.5 per cent modulation.

sary to reduce the flutter to a just detectable value, while the abscissae represent the beat frequency. The noise is 10 db down from an equivalent 30 per cent modulation. The curve is in striking contrast to that of Fig. 5. At very low frequencies a carrier ratio of 28 db is required when a square law detector is employed, while if the receiving set embodies a linear detector a ratio of 1.5 db is sufficient. The right-hand portions of the curves are fairly similar, since the carrier ratio is here dependent upon the audibility of the beat note and not upon flutter effects. The observations of which Fig. 7 is a record were made in the small sound-proof room. In Fig. 8 are shown two curves made in the open laboratory. In the upper curve the noise output was approximately 20 db down from that due to a 30 per cent modulated signal, while in the lower curve it was approximately 30 db down.

The theory which has been outlined indicates that in the case of the square law detector the flutter effects should be practically independent of noise level, and the curves shown in the last three figures bear out this prediction quite positively. Even more definite confirmation is furnished by the upper curve of Fig. 6, which shows the result of observations taken with a fixed beat frequency of 3 cycles. The two curves of this figure show the great superiority of the linear rectifier over the square law in receiving non-isochronous transmissions in the presence of a noise background.

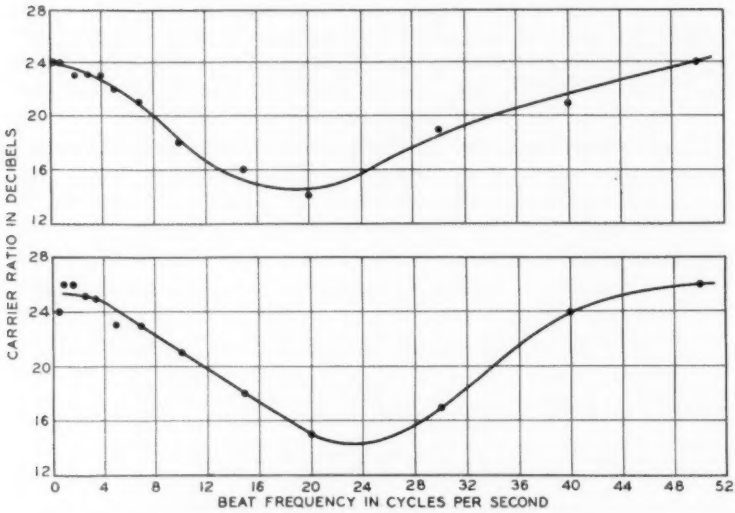


Fig. 8—Carrier ratio for perceptible flutter with a square law detector. Noise equivalent to 3 per cent modulation, for the upper curve, and to 0.95 per cent for the lower curve.

The Square Law Rectifier with Automatic Volume Control

It has been predicted that the use of automatic volume control in the receiving set should greatly increase the flutter effects observable with a linear rectifier, while with a square law device these effects should be the same for both automatic and manual control except, perhaps, at the frequencies of reduced efficiency of the gain control. An experimental check was made on the latter statement, the results of which are shown in Fig. 9. It will be noticed that this curve is very similar to the curves of Figs. 7 and 8.

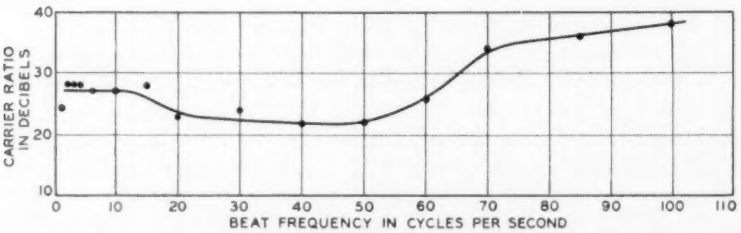


Fig. 9—Carrier ratio for perceptible flutter with automatic volume control. Noise equivalent to 9.5 per cent modulation.

The action of an automatic volume control, in keeping constant the level of the total carrier delivered to the detector, should become less pronounced as the beat frequency rises and should fail altogether when this frequency reaches the audible range. This reduction in efficiency of control may either increase, leave unaltered, or decrease the magnitude of the flutter, depending upon the amount of time delay involved in feeding back the controlling voltage. In the receiver used, the reduction in efficiency of the gain control occurred between 20 and 40 cycles. A comparison of Fig. 9 with Figs. 7 and 8 indicates that in this receiver the gain control tends to increase the flutter somewhat when the heterodyne frequency is within this range.

Interference of Undesired Program

When the interfering station transmits a program which is different from that of the desired station, serious interference may occur which is due primarily to the beats between the undesired sidebands and the desired carrier. If the carrier beat frequency is subaudible and there is little or no noise background, this will be the predominant form of interference. Its magnitude will depend upon the degree of modulation of the undesired signal, but is practically independent of the type of detector and gain control which are used. In the presence of considerable noise background it may or may not be more important than flutter effect.

In order to get some data on this point, observations were made with a square law detector and manual gain control. This represents about the worst condition, as far as flutter effect goes, but will be approximated by AVC receivers. At a fixed noise level the carrier ratio was determined at which the flutter could be noted, and also the ratio at which the program interference was detectable. This was done for receiver band widths of 7000 and 3500 cycles. The band width had no appreciable effect upon the program interference but exercised a very definite effect upon the flutter. Fig. 10 shows the results of the observations which were taken. The solid sloping curve represents the average of the observations on program interference, while the two horizontal curves show the carrier ratio at which the flutter was just detectable for the two bands widths used. The program interference was classed as audible when it could just be heard on the peaks of modulation. However, for considerable intervals of time it was entirely inaudible. Consequently, when the same carrier ratio was recorded for the flutter and for the program interference the former was actually the more annoying. In order to take account of this difference of character between the two types of interference it is necessary to

shift the program curve downward. Just how far it should be displaced is very hard to determine, as the amount will depend upon the type of program on the undesired station. Observations have indicated that the shift should amount to at least 7 db. The dashed curve in Fig. 10 has been drawn 7 db below the solid curve.

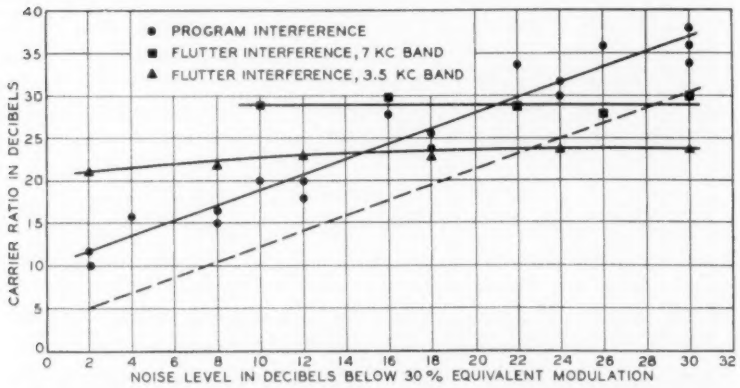


Fig. 10—Relative importance of program interference and flutter interference.

It will be noticed that with a band width of 3500 cycles the flutter curve crosses the program curve at a noise level equivalent to about 2 per cent modulation. (In every case the noise level was measured with the 7000 cycle band, regardless of what band was to be used in the listening tests. This should be kept in mind throughout the present discussion.) This means that at equivalent modulations of more than 2 per cent the flutter effect would be more objectionable than the program interference. However, at high noise levels, say 5 to 10 per cent, the listener would be sure to reduce the band width of his receiver to considerably less than 3500 cycles and this would reduce the relative importance of the flutter. Nevertheless, at very high noise levels the flutter is more important than the program interference. If the undesired station were to employ abnormally low modulation the program interference would be decreased and the relative importance of the flutter increased.

It is evident that the dependence of the flutter on band width, and the different reaction of individual observers as to what type of interference is the more objectionable, renders it impossible to make a definite statement as to the exact values of carrier ratio and noise field which will make the two types of interference equally important. But we can draw the useful conclusion that in cases of excessive noise, such

as may occur in rural areas without causing the listener to abandon attempts at reception, the flutter will be the more important. Consequently, an improvement in the service in such regions would be obtained by synchronizing the carriers of the two stations, even though they continue to transmit different programs.

Effect of Interfering Carrier on Desired Program

Even if the interfering wave were unmodulated and there were a neglibible noise background, there still remains the possibility of distortion of the desired program by the heterodyning action of the undesired carrier. In order to determine how important this effect is as compared with those which have been discussed, a modulated carrier (derived from WABC) and a weaker unmodulated wave were used. A beat frequency of about 3 cycles was maintained during the course of these observations.

With the linear rectifier it was found that a perceptible distortion of the desired program could not be detected on speech and jazz music until the weak carrier was brought within 1 db of the strong one. When the program consisted of music containing many sustained notes, such as occur in a violin solo and even in vocal solos, the cyclic variations in output level were more noticeable. In such a case a ratio of about 4 db was necessary to reduce the distortion to the detectable limit.

With the square law rectifier it was found that a carrier ratio of 10 db produced detectable distortion with any type of program. At a ratio of 16 db distortion could be detected only when the program contained sustained notes, and at 18 db could be noticed only when the notes were sustained for a considerable time.

The dependence of the permissible ratio upon the type of program led us to make a similar observation when the strong carrier was modulated 30 per cent with a pure tone of 400 cycles. Under such conditions it was necessary to reduce the interfering carrier to about 34 db below the strong one before the 3-cycle variation in the pure tone definitely vanished.

Conclusions

The studies which have been reported furnish quantitative data on the various types of interference which are encountered in shared channel broadcasting and show what relative levels of interfering carrier may be tolerated under various conditions.

In high grade service areas the program from the undesired station will be the most serious form of interference, provided the carrier beat

frequency is subaudible. If there is a moderate noise background present, it will tend to mask the program and will therefore permit of somewhat higher interfering field strength. However, if the interference is raised beyond a certain level, dependent upon the received band width, flutter effects will become pronounced. This will not be true with a linear detector and manual gain control, but in practice radio receivers which have linear detectors almost invariably have automatic volume control.

If the noise level is very high it may mask even rather loud program interference, and under such conditions the flutter effect is likely to be much the most serious source of trouble. This condition is of practical occurrence in outlying areas where a degraded service must be tolerated continually. In such regions shared channel broadcasting is limited in usefulness primarily by the flutter effects, and in extreme cases, by distortion of the desired program due to the heterodyning action of the interfering carrier. Both of these types of disturbance would be eliminated by synchronizing the carriers of the two stations, and it seems likely that control of the carrier frequencies to within $\pm$ 0.1 cycle might definitely extend the limits of the lower grade service areas of shared channel stations.

Acknowledgment

I wish to acknowledge my indebtedness to Mr. J. E. Corbin for his assistance in carrying out the experimental work which has been reported in this paper.

Wide-Band Open-Wire Program System *

By H. S. HAMILTON

Radio programs are regularly transmitted between broadcasting stations over wire line facilities furnished by the Bell System. Both cable and open wire facilities are employed for this service. Recently a new program transmission system for use on open wire lines has been developed which has highly satisfactory characteristics. A description of this open wire system and test results obtained with it are given in this paper.

THE simultaneous broadcasting of the same radio program from a large number of broadcasting stations, in different sections of the United States, has become of such everyday occurrence that the radio listening public takes it as an accepted fact and in many cases does not know whether the program is originating in the studio of a local broadcasting station or in a broadcasting studio in some distant city. The wire line facilities furnished by the Bell System for the interconnection of the radio stations, particularly the wire line facilities in cable, have such transmission characteristics that little detectable quality impairment is introduced even when programs are transmitted over very long distances.

This cable program system was described in a recent paper.[1] More recently a new program system for use on open-wire lines, which possesses transmission characteristics comparable with those of the cable system, was developed and an extensive field trial made involving two circuits between Chicago and San Francisco. This paper describes this new open-wire program system and gives the principal results of the tests made on the two transcontinental circuits.

In the paper referred to describing the cable system, the various factors and considerations involved dictating the grade of transmission performance that is desired for program circuits were discussed in considerable detail so they will not be reviewed here. The transmission requirements chosen as objectives for both cable and open wire are as follows:

Frequency Range

Frequency range to be transmitted without material distortion—about 50 to 8,000 cycles.

* Published in April, 1934 issue of *Electrical Engineering*. Scheduled for presentation at Pacific Coast Convention of A. I. E. E., Salt Lake City, Utah, September, 1934.

[1] A. B. Clark and C. W. Green, "Long Distance Cable Circuit for Program Transmission," presented at A. I. E. E. Convention, Toronto, June, 1930; published in *Bell Sys. Tech. Jour.*, July, 1930.

Volume Range

Volume range to be transmitted without distortion or material interference from extraneous line noise—about 40 db which corresponds to an energy range of 10,000 to 1.

Non-Linear and Phase Distortion

Non-linear distortion with different current strengths and phase distortion to be kept at such low values as to have negligible effect on quality of transmission even on the very long circuits.

The frequency range afforded by the new open-wire program circuits extends about 3,000 cycles higher and more than 50 cycles lower than the frequency range available with the open-wire [2] program circuits previously used. The extension of the frequency range at the upper end necessitates the sacrifice of one carrier telephone channel of carrier systems operating on the same wires with the program pair since the frequency band of the lowest carrier channel lies in this range. In order to minimize noise and the possibility of crosstalk, the phantoms of program pairs are not utilized and, of course, d.-c. telegraph compositing equipment is removed in order that the proper low-frequency characteristics may be realized.

Description of New Open-Wire System

In general, the amplifiers on the open-wire program circuits employ the same spacing as the telephone message circuit repeaters on the same pole lead. The average repeater spacing is about 150 miles but the repeaters may be located as close as 60 miles or may be as much as 300 miles apart depending on the location of towns and cities on the open-wire route and the gauge of the wires used. The upper diagram of Fig. 1 shows a typical layout of the new wide-band open-wire program system. Three types of stations are shown, a terminal transmitting station, an intermediate station which may be either bridging or non-bridging and a terminal receiving station.

The terminal transmitting station includes an equalizer for correcting for the attenuation distortion of the local loop from the broadcasting studio, an attenuator for adjusting the transmission level received from the local loop to the proper value, an amplifier for inserting the required gain, filters for separating the program and carrier channels, monitoring amplifier, loudspeaker and volume indicator for

[2] A. B. Clark, "Wire Line Systems for National Broadcasting," presented before the World Engineering Congress at Tokio, Japan, October, 1929; published in *Proc. I. R. E.*, November, 1929, and in *Bell Sys. Tech. Jour.*, January, 1930. F. A. Cowan, "Telephone Circuits for Program Transmission," presented at Regional Meeting of A. I. E. E., Dallas, Texas, May, 1929; published in *Transactions of A. I. E. E.*, Vol. 48, No. 3, pages 1045–1049, July, 1929.

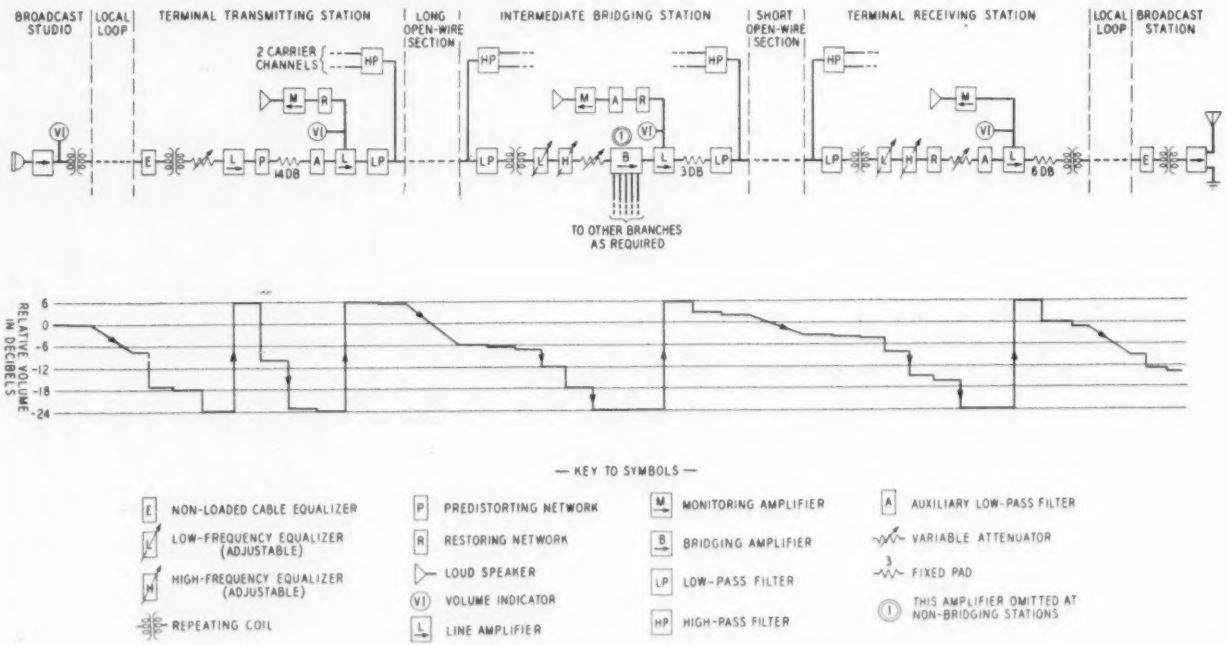


Fig. 1—Typical circuit layout and level diagram.

making the necessary operating observations and a predistorting network and associated amplifier.

At the intermediate station are included line filters for separating the carrier currents and program currents and directing them to their proper channels, two adjustable attenuation equalizers for correcting for the attenuation distortion of the line wires and associated apparatus, gain control attenuator, line amplifier and associated monitoring equipment. At intermediate stations where it is necessary to provide branches to radio stations or to other program circuits an amplifier of a special type having several outlets is inserted immediately in front of the line amplifier.

At a receiving terminal, the layout employed is very similar to that utilized at intermediate stations except that an additional low-pass filter and a restoring network are inserted ahead of the receiving amplifier.

A novel feature is provided in this program system for minimizing its susceptibility to interference at higher frequencies. It consists in predistorting the transmission at the sending end of the circuit so that currents above 1,000 cycles are sent over the line at a higher level than if this arrangement were not employed, thus increasing the signal-to-noise ratio at these frequencies. Such an increase in power at high frequencies is permissible without overloading in the line amplifiers in view of the fact that the energy content of the program material above 1,000 cycles is materially less than at the low frequencies and decreases rapidly as the frequency is increased. In order to restore the program material to the same relations it would have if it were not predistorted, a network is inserted at each point in the branches which feed the radio stations and at the receiving terminal. This network introduces attenuation and phase distortion which are complementary to those introduced at the sending end of the circuit by the predistorting network. The net reduction in high-frequency interference is equal to the discrimination introduced by the predistorting network in favor of these frequencies, and is therefore equal approximately to the loss of the restoring network at the same frequencies.

In the lower part of Fig. 1 is shown a level diagram, from which may be noted the losses and gains introduced by different parts of the system at a frequency of 1,000 cycles. The maximum volumes which are permitted in the various parts of the system are also indicated approximately by this diagram.

Line Facilities

As is well known the open-wire lines employed in telephone and program service do not have uniform attenuation for all frequencies,

the low frequencies being transmitted with much less loss than the high frequencies. Since the program circuits employ the same type of open-wire facilities that is used in the message circuits, three different gauges of wire with either of two pin spacings between wires may be used and the repeater sections may vary in length from 60 to 300 miles. This means that the attenuation frequency characteristic of a repeater section not only varies with frequency but also varies considerably in magnitude of attenuation depending on gauge of wire and length of repeater section.

On Fig. 2 are shown three pairs of characteristics which illustrate the loss-frequency characteristics of three lengths of 165-mil, 8-inch

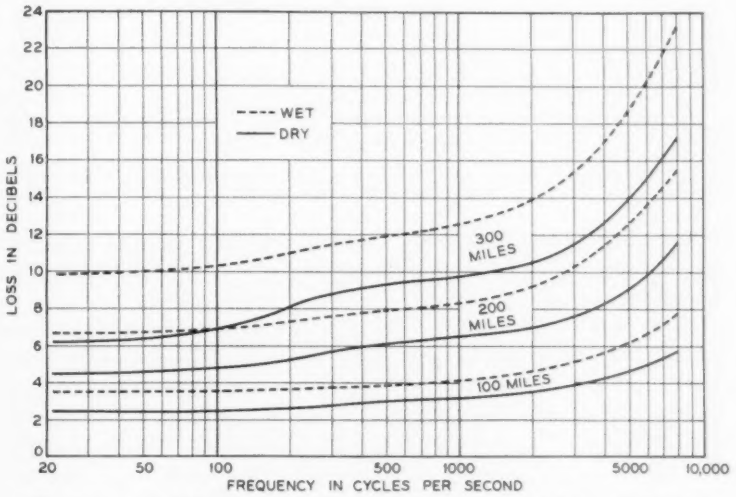


Fig. 2—Loss of 165-mil. 8-inch spaced pairs when inserted between 600-ohm resistances.

spaced circuits. The lengths chosen for purposes of illustration are 100, 200 and 300 miles, respectively. The solid line curves show the insertion loss-frequency characteristics of the circuits for average dry weather conditions when the circuits are connected between 600-ohm resistances. The dashed line curves indicate the wet weather insertion loss characteristics, that is, they indicate the loss-frequency characteristic which might obtain if the lines were very wet for the entire length of a repeater section.

For the purpose of comparing the attenuation frequency characteristics of the different types of open-wire lines, the curves shown on Fig. 3 have been prepared. These characteristics have been plotted so

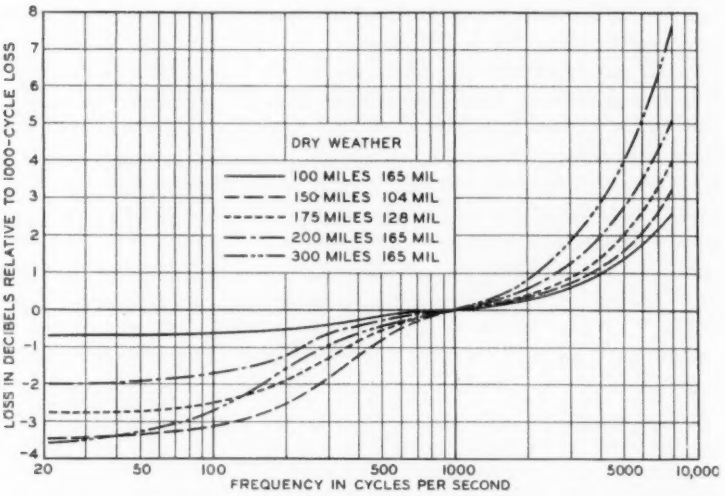


Fig. 3—Attenuation characteristics of 8-inch spaced open-wire pairs when inserted between 600-ohm resistances.

that all coincide at 1,000 cycles; thus a direct comparison of the difference in shape of the attenuation frequency characteristics may readily be observed.

Figure 4 shows resistance and reactance components of 165-mil and 128-mil 8-inch spaced open-wire lines. Note that, except at low frequencies, the impedances of the various open-wire lines are quite uniform throughout the frequency range and do not depart greatly from 600 ohms. For this reason and in consideration that the majority of telephone apparatus is designed for 600-ohm impedance, all units of this new program system, except the carrier line filters, have been designed to have an impedance of 600 ohms. In order to reduce reflection losses, particularly in the carrier range, the line filters have been designed to have an impedance on the line side somewhat lower than 600 ohms although the drop or office side impedance is 600 ohms.

Attenuation Equalizers

To furnish the necessary attenuation corrections for the three different gauges of lines, four adjustable attenuation correcting networks have been provided. One attenuation equalizer provides attenuation correction for high frequencies only and is common for all gauges. The three other equalizers provide low-frequency attenuation correction designed specifically for the particular gauge of circuit the

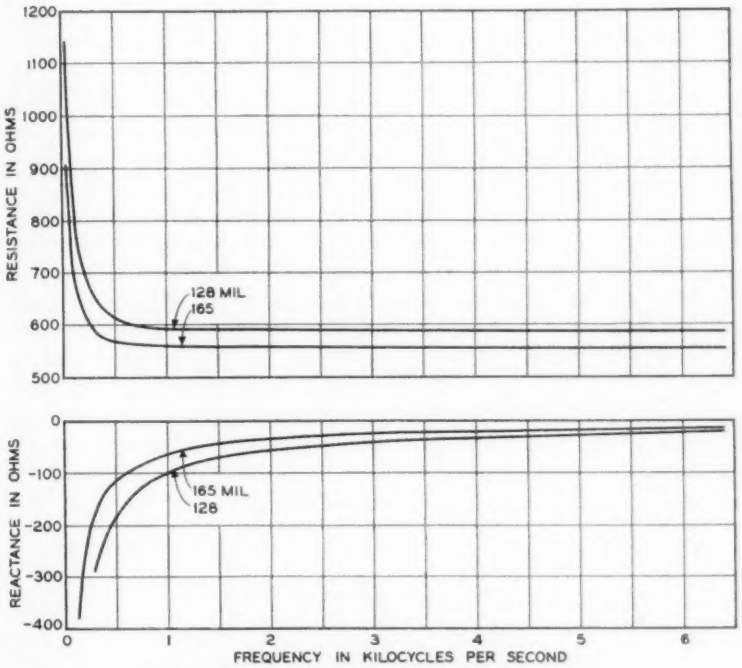


Fig. 4—Impedance of 8-inch spaced open-wire pairs.

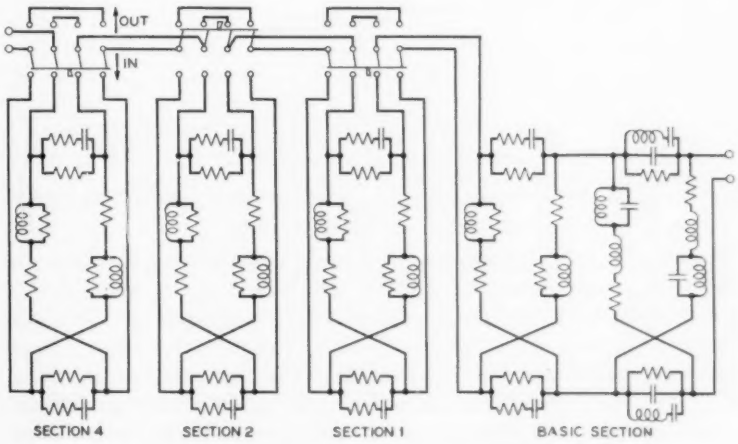


Fig. 5—Low-frequency attenuation equalizer.

equalizer is to be associated with and also include a fixed amount of high-frequency attenuation correction.

On Fig. 5 is shown a schematic diagram of one of the low-frequency attenuation equalizers. This consists of four sections of 600-ohm constant impedance type networks. One section referred to as a basic section introduces attenuation correction over the complete frequency range from 35 to 8,000 cycles for a particular minimum length of line, as for example, in the case of 165-mil circuits this is for 100 miles. The three other sections on the other hand furnish attenuation correction only for frequencies from approximately 1,000 cycles down to 35 cycles. Section 1 of the equalizer for 165-mil circuits puts in about ½ db more loss at low frequencies than it does at 1,000 cycles. Section 2 puts in double the amount of correction that is introduced by Section 1 and Section 4 introduces four times as much attenuation correction as Section 1. These three sections are controlled by switches so that any one or all of them may be cut in tandem with the basic section. The attenuation corrections afforded for the various adjustments of this equalizer are shown on Fig. 6.

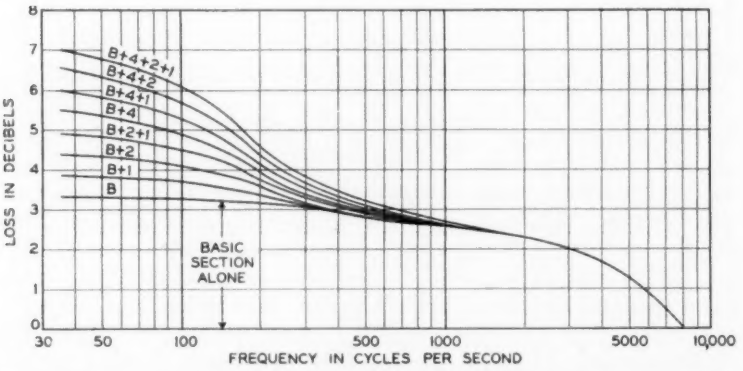


Fig. 6—Attenuation correction furnished by low-frequency equalizer for 165-mil. circuits.

The attenuation equalizers for 128-mil and 104-mil facilities are similar in construction to the one just described having different constants so as to furnish somewhat different attenuation correcting characteristics.

Figure 7 shows a schematic diagram of the high-frequency attenuation equalizer. This consists of four 600-ohm constant impedance type network sections which, as indicated, are controlled by switches so that any one or all of them may be cut in tandem with the program

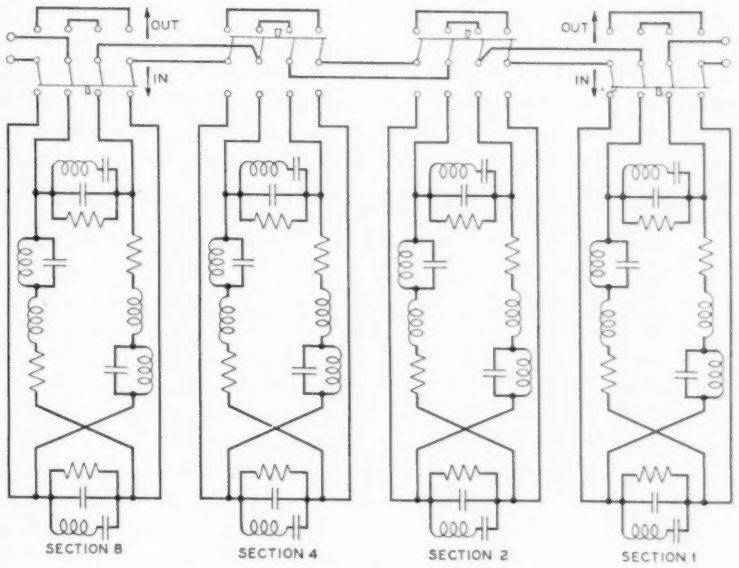


Fig. 7—High-frequency attenuation equalizer.

circuit as required. Figure 8 shows the loss-frequency characteristics of these four sections. As may be noted, the loss of the various sections is practically constant over the frequency range up to 1,000 cycles, decreasing from there on to a minimum value at 8,000 cycles.

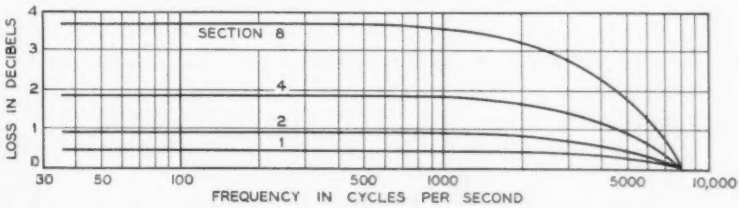


Fig. 8—Attenuation correction furnished by high-frequency equalizer.

Section 1, as may be noted, furnishes about ½ db attenuation correction. Section 2 is double that of Section 1, Section 4 is four times that of Section 1 and Section 8, eight times that of Section 1. These sections may be used in tandem so that attenuation correction for the high frequencies is, therefore, provided in steps of ½ db from zero to 7½ db.

An illustration of how the equalizers introduce the necessary attenuation correction is given on Fig. 9. The lower curve on this figure shows the loss of a 300-mile section of 165-mil circuit. The losses introduced by the particular sections of low and high-frequency equalizers that

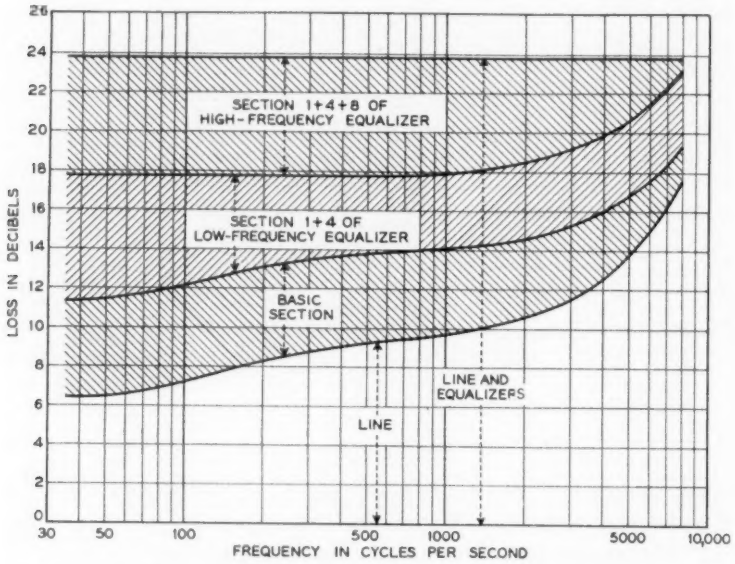


Fig. 9—Loss of 300-mile line section and associated equalizers.

would be required for this length of line are indicated by the cross-hatched areas, and the total line and equalizer loss is shown by the top horizontal line. Sufficient gain is introduced by the line amplifier to annul this loss.

Amplifiers

Two types of amplifiers are provided, one of which is used as a line or monitoring amplifier and the other which is used as a means for transforming one circuit into several circuits so as to feed various branches at points required.

For certain combinations of program circuits as many as 50 amplifiers may be connected in tandem. This necessarily imposes severe requirements on the transmission performances of the amplifiers particularly with reference to flatness of gain-frequency characteristics and phase distortion. By designing the coils used in the amplifiers

so as to have very high inductances the desired phase distortion requirements were met while at the same time the necessary flatness of gain characteristic was obtained at the low frequencies.

Figure 10 shows the transmission circuit of the line amplifier and monitoring amplifier. This device has a 600-ohm input and output impedance and consists of two stages of push-pull amplification. The potentiometer is a balanced slide wire having a continuous gain adjustment over a range of 6 db. A balanced input transformer serves to connect the potentiometer to the grids of the two push-pull vacuum tubes which function as the first stage of this amplifier. The first stage is connected to the second or power stage by means of resistance coupling which gives better results both as to phase distortion and low-frequency gain characteristics than if transformer or retard coil coupling were used. Resistances are provided in the grid circuits of the second stage so that the high-frequency characteristic may be adjusted as required. The power tubes are connected to an output transformer which has the unique feature of providing a monitoring outlet which is not materially affected by voltages produced at or beyond the line terminals. The transformer, as may be observed, consists of three balanced windings arranged as in the form of the well-known hybrid coil used in two-wire telephone repeaters, with the exception that the two low impedance windings are of unequal ratio, the line windings having many more turns than the monitoring windings. The ratio of the windings is such that the voltage at the monitoring terminals when said terminals are closed through 600 ohms is 30 db below the voltage at the line terminals. Resistances are inserted in series with the monitoring winding so that an impedance of 600 ohms will be presented at the monitoring terminals.

The average gain of the amplifier with the potentiometer set at its maximum position is 33 db. Of 100 amplifiers measured, the gain at 35 cycles averaged .10 db less than the gain at 1,000 cycles while from 100 to 8,000 cycles the gain was constant within .05 db. The delay at 50 cycles is approximately .6 millisecond greater than it is at 1,000 cycles. From 150 to 8,000 cycles the delay is substantially constant and is only a small fraction of a millisecond. The amplifier is capable of handling an output power 9 db above reference volume without noticeable distortion.

At several points along a program circuit taps or branches are provided so as to connect various broadcasting stations to the program circuit and also to connect to other program circuits which form part of a broadcasting network. Points where such connections or branches are made are commonly called bridging stations. At some points as

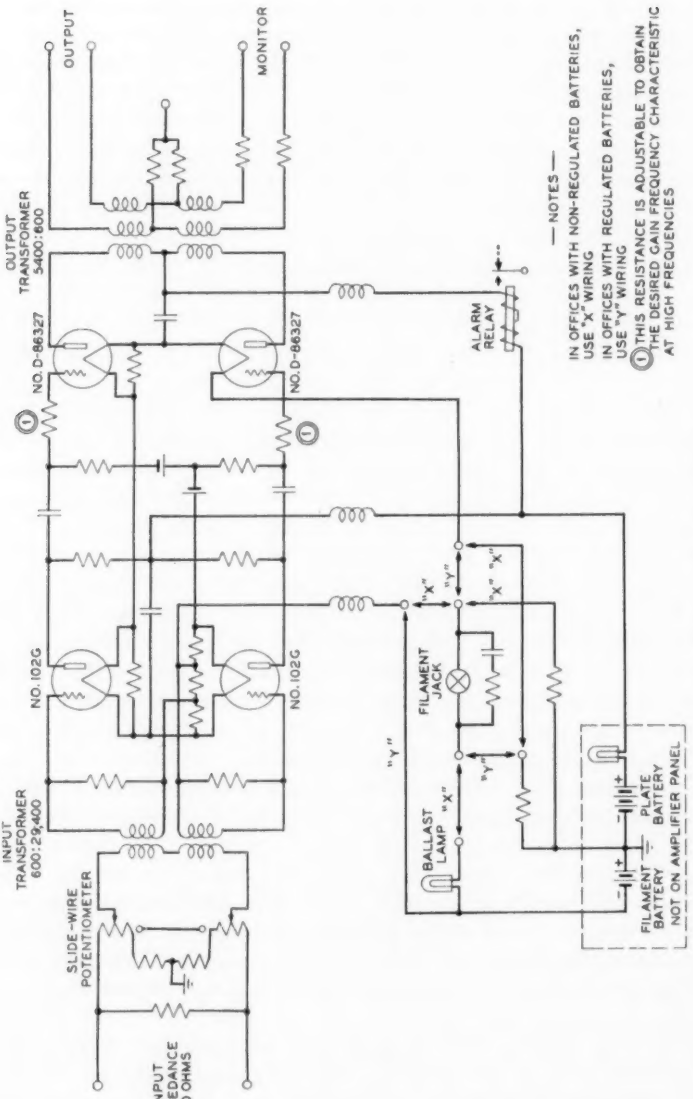


Fig. 10—Schematic of line amplifier.

many as six branches are supplied but generally only two or three taps are utilized.

To accomplish this branching out at a bridging station a resistance network multiple is provided, having six outlets. This network multiple is shown on Fig. 11. To annul the loss of the network a single stage amplifier is connected in front of it. This network multiple and amplifier are mounted on the same panel forming a single integral unit. The network multiple is so proportioned that if any one of the branches is accidentally opened or short-circuited the other branches are affected to only a minor degree. The amplifier is adjusted so that the gain from the input terminals to any of the output branches is zero. The bridging amplifier is normally inserted immediately in front of the line amplifier. As in the case of the line amplifier mentioned above, high inductance coils are utilized in order to keep phase distortion at a minimum. A resistance adjustment is provided in the grid circuit in order to adjust the high-frequency characteristic of this amplifier to the desired value.

The gain-frequency characteristic of the bridging amplifier is practically identical with the corresponding characteristic just described for the line amplifier, while the delay is even less.

Predistortion

The means utilized to accomplish the predistorted transmission referred to earlier includes the provision of a so-called predistorting network at the sending end of a program circuit and a restoring network in each branch which supplies a broadcasting station. The predistorting network introduces a large loss at low frequencies with a decrease in loss as the frequency is increased. By introducing suitable amplification immediately behind the predistorting network the resultant effect is to raise the high-frequency transmission relative to the low-frequency transmission by the difference in loss between the 1,000-cycle loss of the predistorting network and its higher frequency loss. The restoring network characteristic is the inverse of the predistorting network. These two networks are 600-ohm constant impedance type structures. The restoring network is shown schematically in Fig. 12. The predistorting network is generally similar to this, having different constants and a slightly different arrangement of elements. On Fig. 13 are shown the loss-frequency characteristics of the predistorting and restoring networks and a third characteristic which is the sum of these two. As may be noted this latter characteristic has a constant value throughout the frequency range.

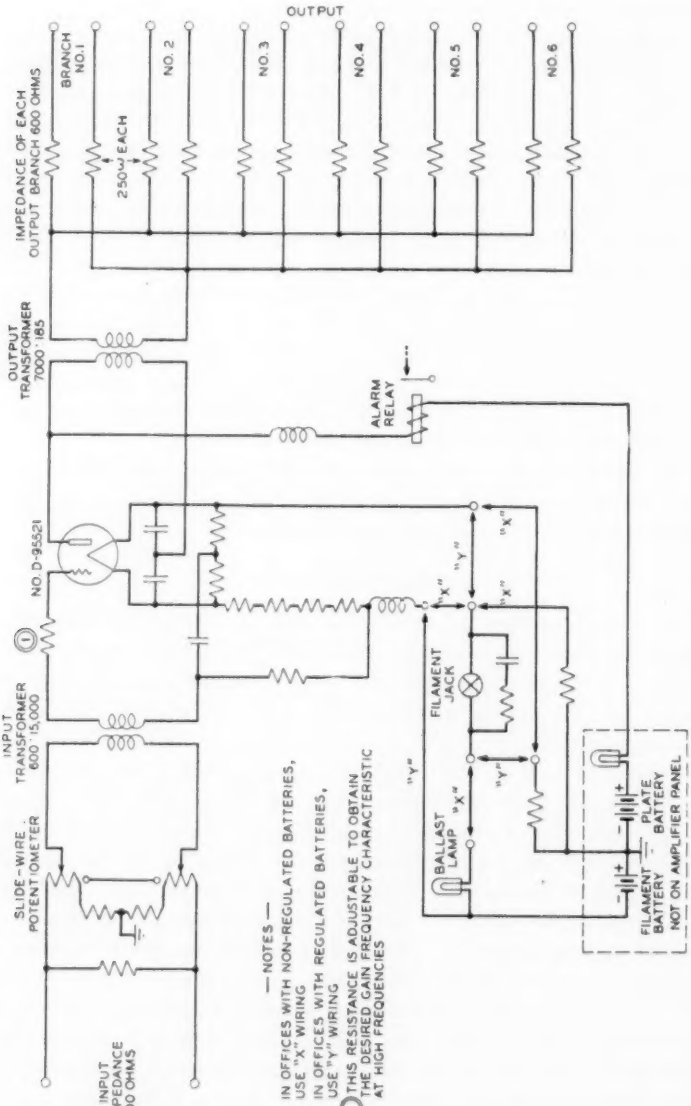


Fig. 11—Schematic of bridging amplifier.

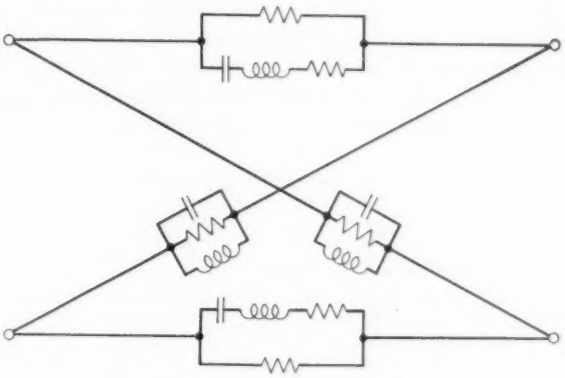


Fig. 12—Restoring network.

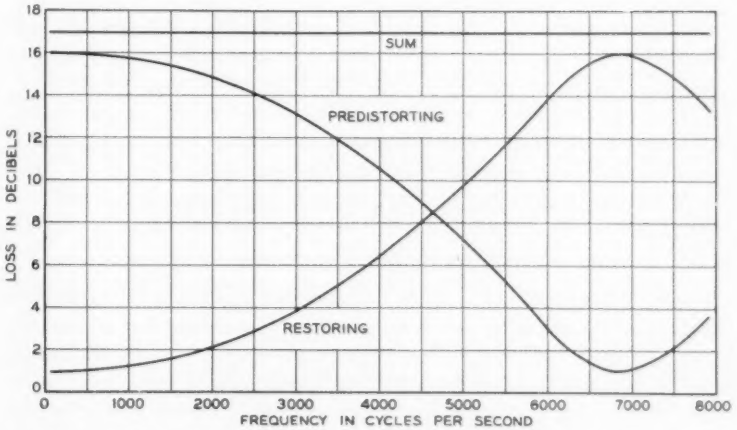


Fig. 13—Attenuation characteristics of predistorting and restoring networks.

Line Filters

As a rule, on open-wire circuits other transmission channels are provided on the same wires which carry the program transmission. These other channels operate at frequencies above the program range and in order to direct the various currents to their proper channels at a terminal or repeater station carrier line filter sets are inserted at the ends of the line wires. The carrier line filter sets include a low-pass and a high-pass filter. The low-pass filter, cutting off somewhat above 8,000 cycles, directs the program transmission to the program

apparatus and the high-pass filter which has a low end cutoff around 9,000 cycles directs the carrier transmission to its associated carrier equipment. Attenuation frequency characteristics of these filters are shown on Fig. 14. The low pass filter is of unusual design and is described at some length in a companion paper.[3]

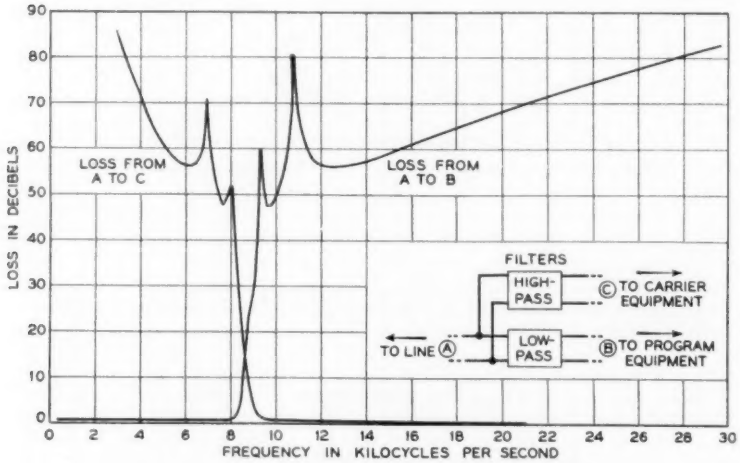


Fig. 14—Attenuation characteristics of line filter set.

Monitoring Features

A very important factor in the satisfactory operation of a program system is the provision of monitoring arrangements by which the operating forces are enabled to observe the quality of transmission, listen for extraneous interferences and observe indicating devices in order to make certain that the program is maintained at its proper volume.

Three types of aural monitoring facilities were provided on a trial basis for the new program system. The first type consists of a single unit loudspeaker operated by a suitable amplifier. With this loudspeaker system a good response characteristic from approximately 100 to 5,000 cycles is obtained, the low-frequency response depending, of course, on the size of the baffle used with the loudspeaker.

The second type of monitoring consists of two headset receivers arranged with a proper equalizing network circuit. This type of monitoring provides good response characteristics from approxi-

[3] A. W. Clement, "Line Filter for Wide-Band Open-Wire Program System," published in this issue of the *Bell Sys. Tech. Jour.*

mately 50 cycles to 8,000 cycles, enabling the observer to cover the entire program frequency range, thus permitting him to detect any extraneous interference which may be introduced even though this occurs at very low or very high frequencies.

The third type of monitoring consists of two loudspeakers and associated equalizing network with the loudspeakers mounted in a large baffle board. This arrangement affords a fairly uniform response from about 40 cycles to above 8,000 cycles. The particular type of monitoring which might be provided at the various stations would be governed by the service requirements involved.

To observe the volume on the program circuit, volume indicators are used. A new type of volume indicator was made available along with the new program system. This new device utilizes a full-wave copper oxide rectifier, has a much greater sensitivity range than that of the devices formerly used and possesses materially improved indicating characteristics. The volume indicator is connected across the monitoring terminals of the line amplifier, in which position it is bridged across a practically non-reactive 600-ohm impedance. Located thus it is also independent of line impedance affording more accurate results and obviating the necessity of correcting volume readings on account of line impedances. Also at this location it introduces no loss or phase distortion to the through program circuit.

The above constitutes a description of the major items employed in this program system. There are a number of other units, such as attenuators, repeating coils, etc., which will not be described in detail here but will be referred to as the need arises.

Typical Station Layouts

Due to the various requirements for different types of service and due in part to the different type of facilities, the general apparatus layouts and arrangements at different repeater stations are not always the same. Several of the more important general or typical layouts will be briefly discussed, however.

On Fig. 15 is shown a layout of a typical intermediate station where bridging is not required and where the gauge of the wires in the two directions is the same. As may be noted from this figure, switching facilities are provided so that the apparatus may be connected into the circuit so as to properly take care of either the east-west or west-east transmission. For this type of layout most of the apparatus is common to both directions of transmission. The fixed artificial lines or pads indicated by Note 1 on Fig. 15 are for the purpose of building out whichever line has the lower 1,000-cycle attenuation so that this

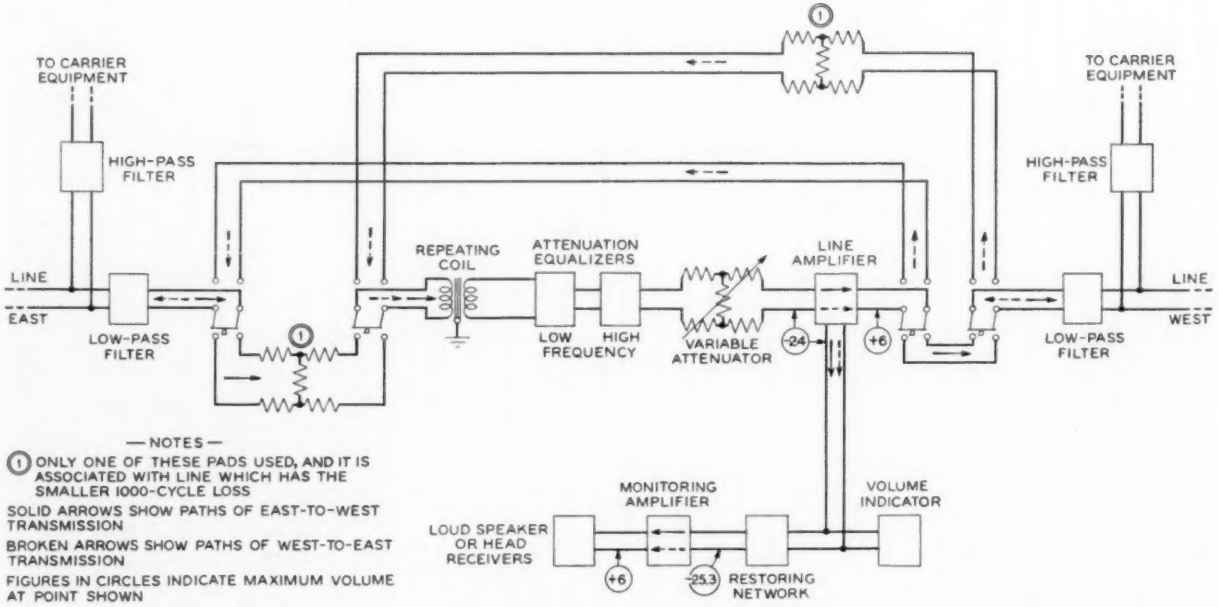


Fig. 15—Intermediate station layout.

line and associated pad will have the same 1,000-cycle loss as the other line. As indicated, only one of these pads is required. This building out of the shorter line minimizes attenuator adjustment when the direction of transmission is reversed. The line amplifier in this, as well as the other layouts to be discussed, is always set for a gain of 30 db.

On Fig. 16 is shown the layout of a typical intermediate non-bridging station where the gauges of the wires on the two sides of the repeater station are different. As mentioned earlier each gauge of wire has its own particular low-frequency attenuation equalizer. Consequently, where the gauges of the wires on the two sides of the repeater station are not alike, it is necessary to arrange the station layout so that the proper low-frequency equalizer will be associated with the proper direction of transmission. This association of apparatus may be readily observed from Fig. 16.

On Fig. 17 is shown the layout of a typical terminal station. This layout differs from the intermediate station layout largely in the fact that provision must be made for the introduction of predistortion when the terminal station is transmitting a program to the open-wire line and in the provision of a restoring network when the terminal station is receiving a program from the open-wire line. The general layout of the apparatus may readily be observed by reference to the figure. The monitoring facilities at this type of station, in general, differ from those provided at the normal intermediate station in that a two-unit loudspeaker is provided for use as desired.

On Fig. 18 is shown the layout of a typical intermediate bridging station where the gauge of the wires in the two directions is the same. This arrangement differs largely from the arrangement shown on Fig. 15 in that the bridging amplifier is inserted immediately ahead of the line amplifier so as to provide the necessary additional branches as required. The general circuit arrangements involved to take care of the different types of branches which may be encountered are indicated on this figure. The photograph, Fig. 19, shows the program equipment layout at an intermediate bridging station, which is of the type just discussed in Fig. 18, utilizing, however, only one branch circuit which is connected to a local broadcasting station.

In certain of the layouts just discussed, one apparatus unit designated as "Aux Filter" is shown which has not previously been mentioned. This is an 8,000-cycle low-pass filter somewhat similar to the low-pass line filter, except that it is not designed to operate in parallel with any high-pass filter. This filter is required at the transmitting and receiving terminals, in the branches feeding the radio station and

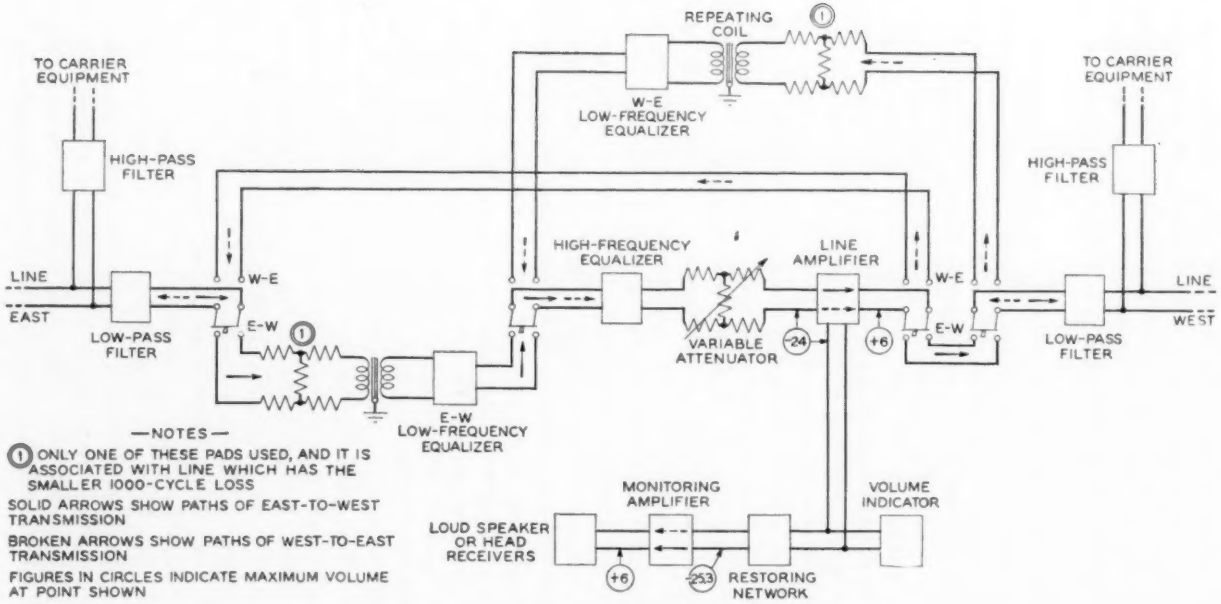


Fig. 16—Intermediate station layout; different gauges.

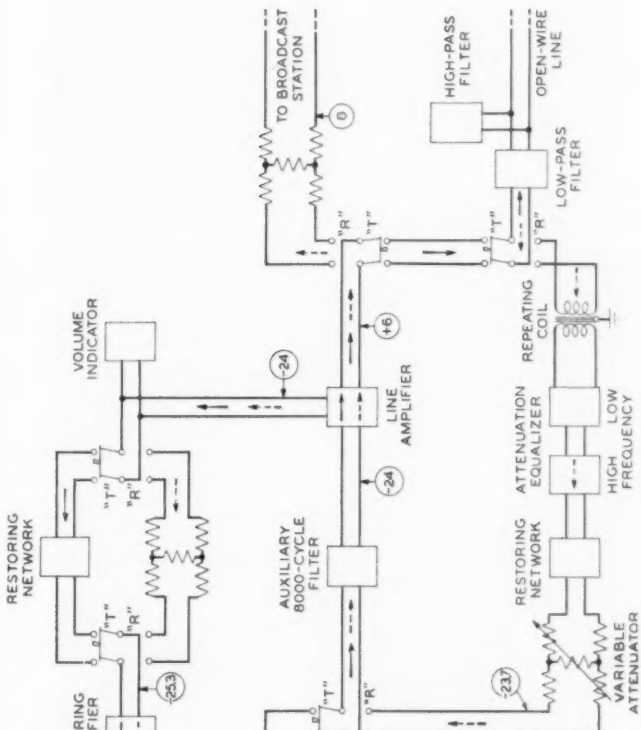


Fig. 17—Terminal station layout.

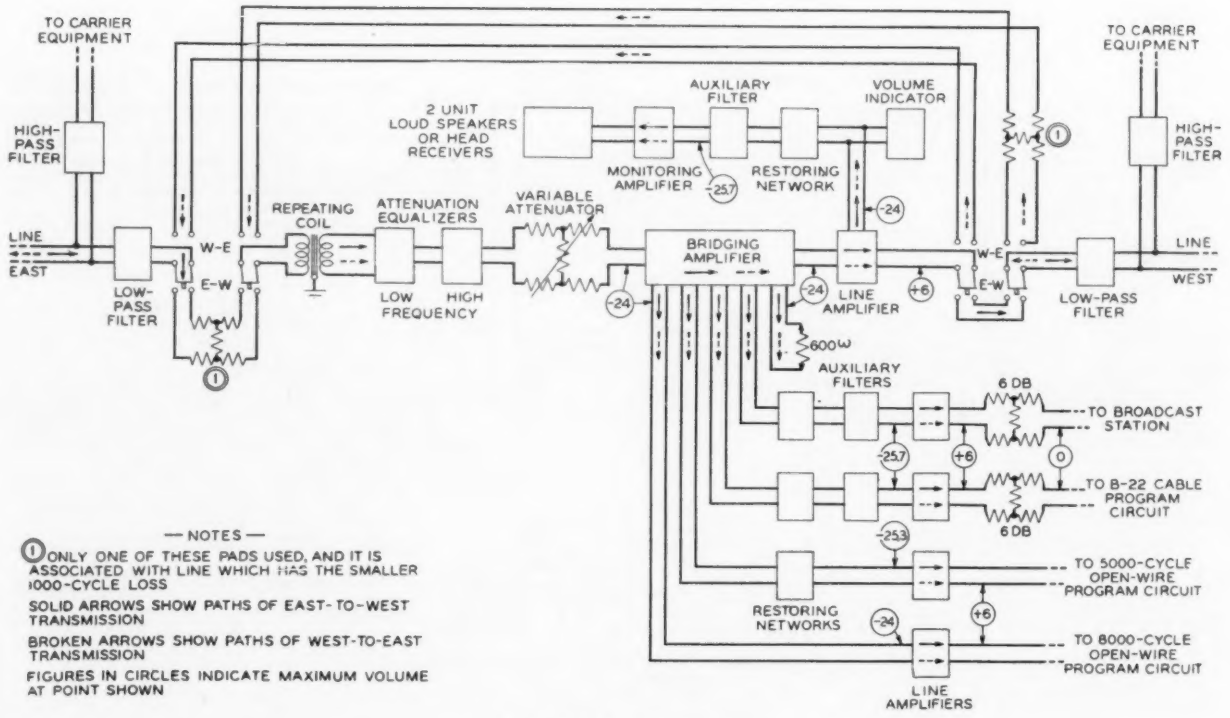


Fig. 18—Bridging station layout.

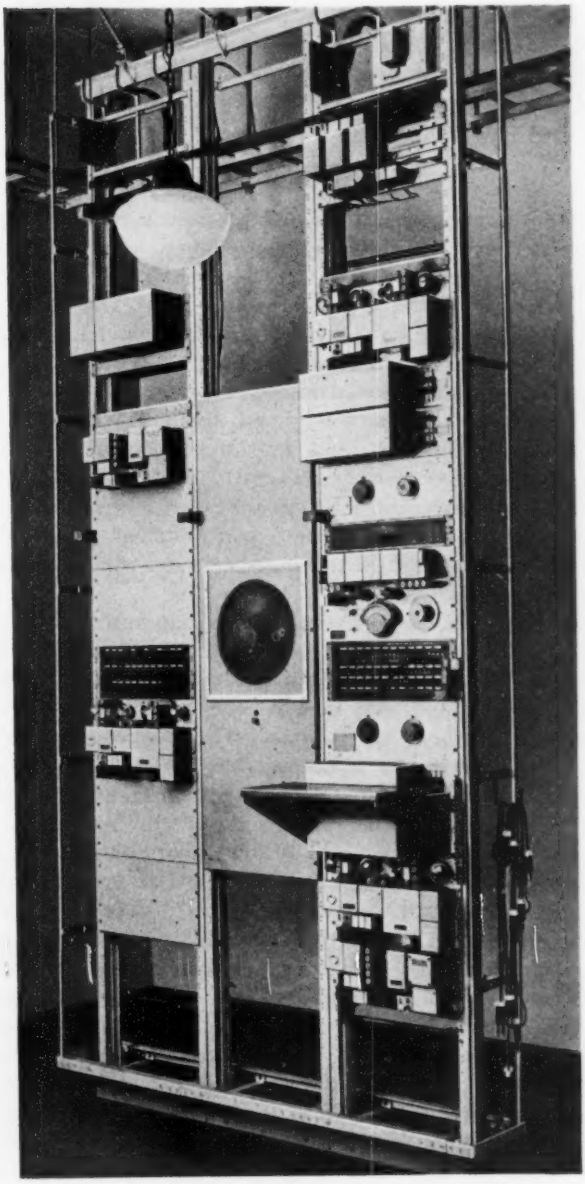


FIG. 19—Intermediate bridging station bay layout.

also in the high quality monitoring circuit to afford additional discrimination against unwanted high-frequency interference as, for example, interference from the carrier channels. This arrangement of splitting the filter requirements enables a less expensive type of line filter set to be employed.

Overall Performance

The initial application of this new program system was made on two transcontinental circuits between Chicago and San Francisco. One circuit, referred to as circuit 1, was routed through Omaha and Denver over the central transcontinental line. The other circuit, referred to as circuit 2, was routed via St. Louis and Kansas City to Denver and thence over the same pole lead as circuit 1. The layout of these two circuits is shown in Fig. 20. Circuit 1 was approximately 2,395 miles long and was routed through 17 repeater stations involving 23 amplifiers in tandem. Circuit 2 was approximately 2,689 miles long and was routed through 19 repeater stations involving 29 amplifiers in tandem. Both circuits were routed through B-22 cable facilities between Sacramento and Oakland, California, and non-loaded cable facilities in the transbay submarine cable between Oakland and San Francisco.

At San Francisco a listening studio was set up in the Grant Avenue office where the program circuits terminated. A two-unit loudspeaker with suitable connecting networks was set in a 7′ x 7′ baffle, the response of this loudspeaking system being practically uniform from about 40 cycles to above 8,000 cycles. The room in which the loudspeakers were located was acoustically treated so as to obtain the proper reverberation time. A powerful amplifier having a flat gain-frequency characteristic from 35 cycles to well above 8,000 cycles supplied the loudspeaker system. A high quality phonograph system for furnishing test programs was also installed at the Grant Avenue office. The records used were of the vertical cut type and included several recordings of a 75-piece orchestra as well as various solo and instrumental recordings. Two outside pickup points were used, one at the studios of one of the broadcasting companies at San Francisco and the other at a hotel. At both of these places the moving coil type of microphones was used and the latest type of high quality pickup amplifiers. The pickup system used at both these places had a response characteristic within about 2 db of being flat over the range of 35 to 10,000 cycles.

Figure 21 is a photograph showing the special equipment placed in the Grant Avenue office for carrying out the various overall tests and

TO DENVER

TO DENVER

129 M

165

12 IN

TO DENVER

176

165

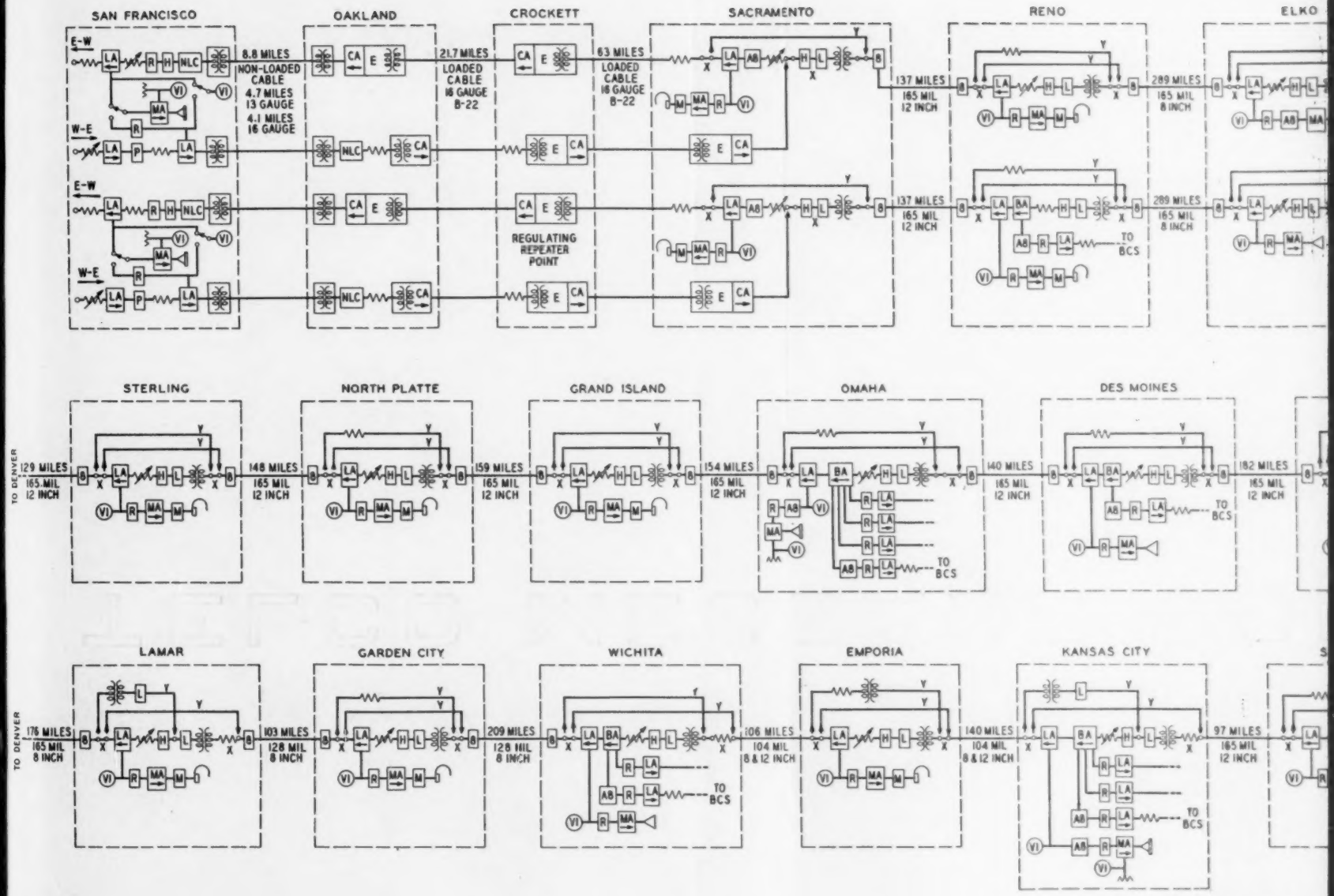


Fig. 20—Circuit layout for trial of wide-band

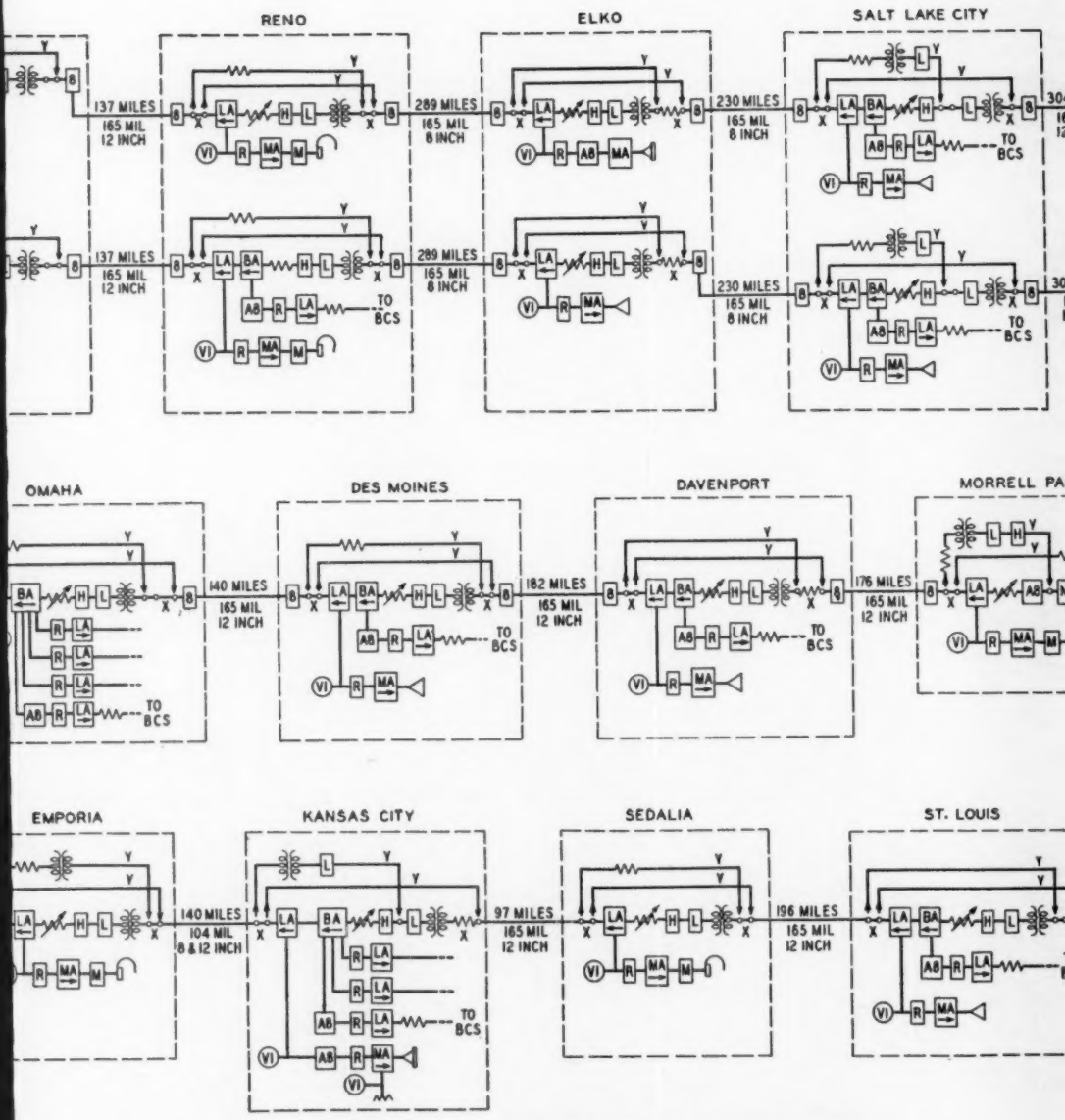
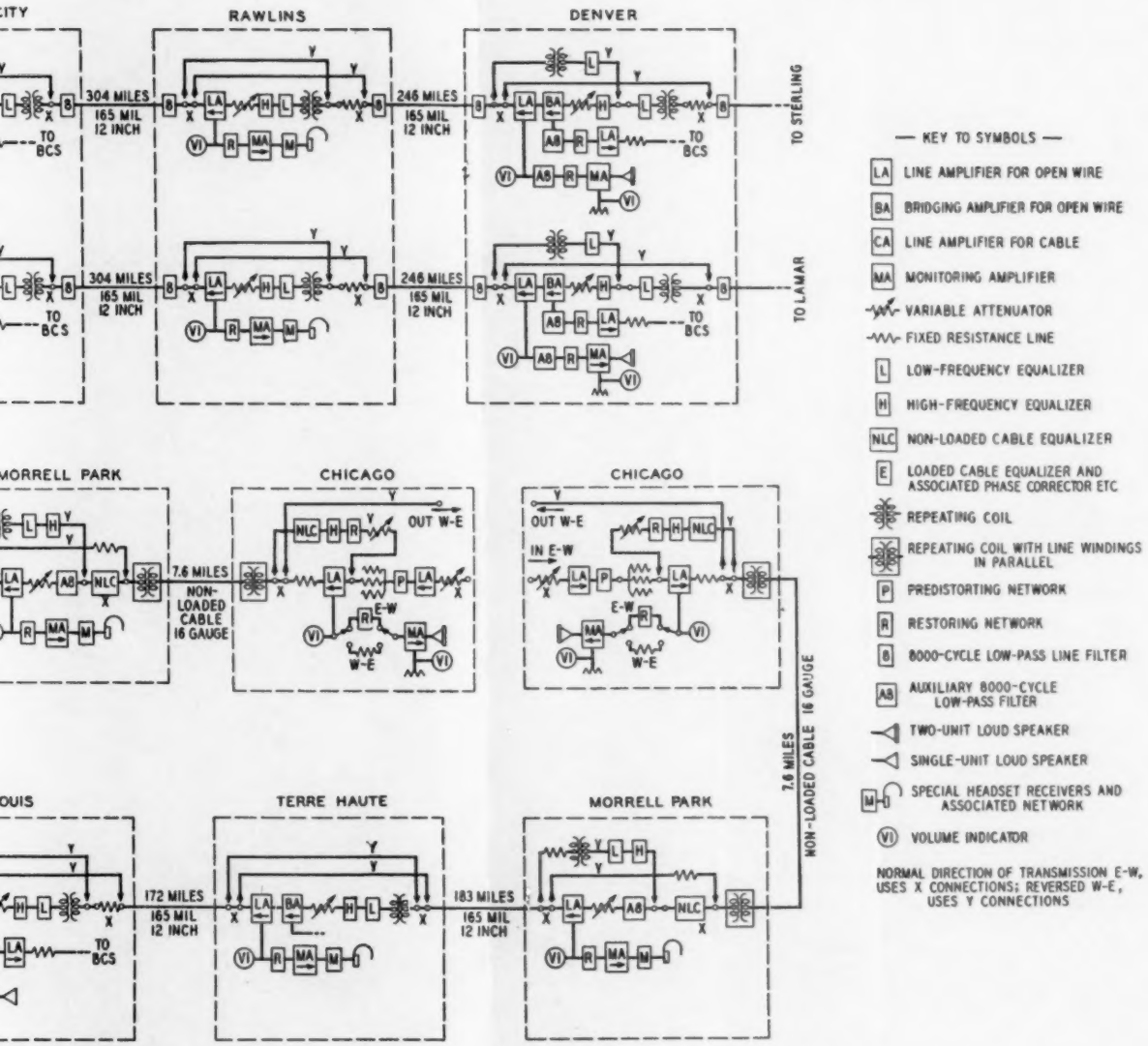


Fig. 20—Circuit layout for trial of wide-band open-wire program system.



— KEY TO SYMBOLS —

- LA — LINE AMPLIFIER FOR OPEN WIRE
- BA — BRIDGING AMPLIFIER FOR OPEN WIRE
- CA — LINE AMPLIFIER FOR CABLE
- MA — MONITORING AMPLIFIER
- VARIABLE ATTENUATOR
- FIXED RESISTANCE LINE
- L — LOW-FREQUENCY EQUALIZER
- H — HIGH-FREQUENCY EQUALIZER
- NLC — NON-LOADED CABLE EQUALIZER
- E — LOADED CABLE EQUALIZER AND ASSOCIATED PHASE CORRECTOR ETC
- REPEATING COIL
- REPEATING COIL WITH LINE WINDINGS IN PARALLEL
- P — PREDISTORTING NETWORK
- R — RESTORING NETWORK
- B — 8000-CYCLE LOW-PASS LINE FILTER
- AB — AUXILIARY 8000-CYCLE LOW-PASS FILTER
- TWO-UNIT LOUD SPEAKER
- SINGLE-UNIT LOUD SPEAKER
- M — SPECIAL HEADSET RECEIVERS AND ASSOCIATED NETWORK
- VI — VOLUME INDICATOR

NORMAL DIRECTION OF TRANSMISSION E-W, USES X CONNECTIONS; REVERSED W-E, USES Y CONNECTIONS

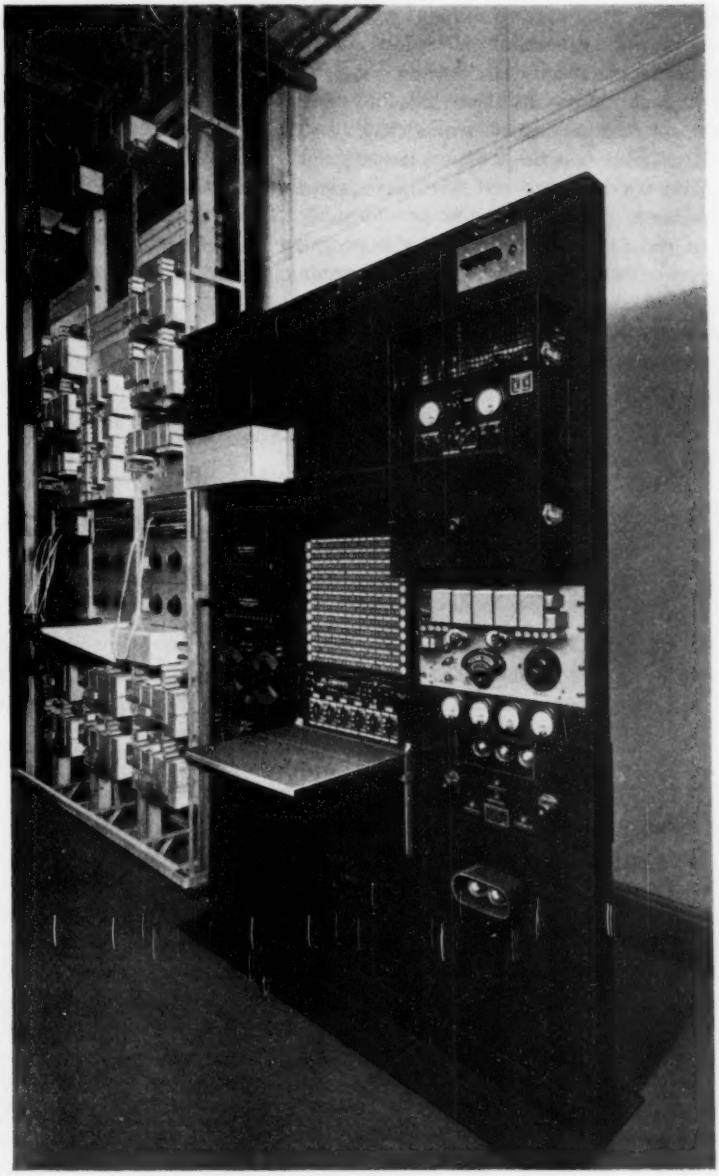


Fig. 21—Special apparatus bay layout.

also shows the new equipment provided at San Francisco on the two program circuits under discussion. The three right-hand bays accommodated the special equipment.

In making transmission measurements, the circuit under test was first split up in a number of sections and each section was then measured at four test frequencies, namely, 50, 100, 1,000 and 7,000 cycles. If the results were not within required limits the attenuators and equalizers were readjusted as required. The various sections were then connected together and the overall circuit measured at several frequencies. Figure 22 shows the transmission-frequency character-

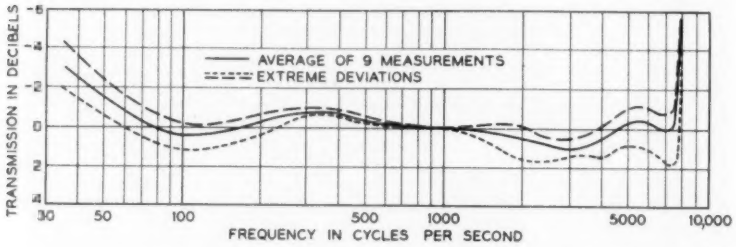


Fig. 22—Transmission frequency characteristics of circuit No. 1, Chicago to San Francisco.

istics of circuit 1. The solid line is the average of nine measurements while the dashed lines show the extreme deviations obtained for any of the nine measurements. Figure 23 shows corresponding data for

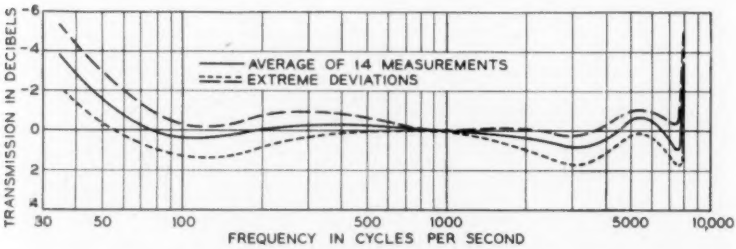


Fig. 23—Transmission frequency characteristics of circuit No. 2, San Francisco to Chicago.

circuit 2. For comparison purposes the average characteristics of the two circuits separately and the two of them connected in tandem making a loop circuit of over 5,000 miles are shown on Fig. 24.

Other measurements were made to determine whether non-linear effects were produced. For example, two frequencies were applied

to the circuit, one being measured and the other alternately cut off and on to determine whether one frequency adversely affected the transmission of the other or produced undesirable sum and difference products. Such distortion effects were found to be small. Measurements were made to determine whether the overall transmission varied with the load applied. With a testing power which was varied in magnitude from 50 milliwatts to .1 milliwatt, the transmission varied slightly more than 1 db, that is, with the heavy load the circuit loss was somewhat more than 1 db greater than at the light load.

A noise and crosstalk survey was made on these program circuits and on message circuits on the same pole lead. Observations were made at the terminals of the message circuits while a program was being transmitted on the program circuits to determine the amount of

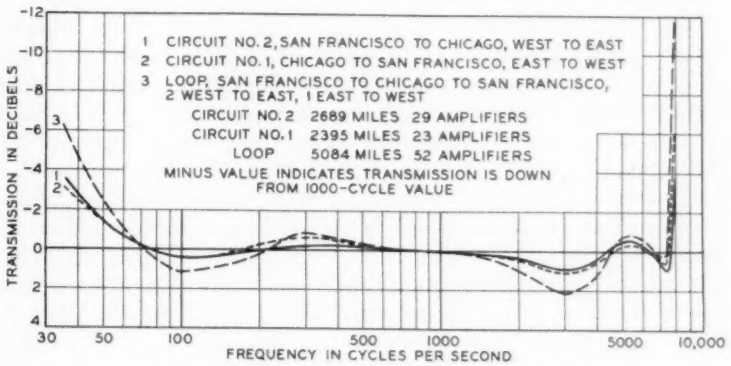


Fig. 24—Average transmission frequency characteristics.

interference introduced into the message circuits from the program circuits, and, conversely, observations were made on the program circuits while various paralleling message circuits were in use, and the resulting interference was recorded.

The noise or crosstalk volume on the program circuits was measured by means of a volume indicator, which had inserted between it and the circuit at the point of measurement a network having a loss-frequency characteristic such that the various frequencies affecting the meter reading were attenuated or weighted in much the same way that the ear weights the different frequencies. Crosstalk volume and noise on the message circuit were measured with an indicating meter in much the same manner except that the network used here had an attenuation frequency characteristic corresponding very nearly to that of the ear and an average telephone set. The network used on the program

circuits was referred to as a "program weighting network," while that used with the message circuit was the ordinary "message weighting network." The noise and crosstalk volume was then recorded in db referring to reference noise with either program weighting or message weighting. Reference noise is that amount of interference which will produce the same meter reading as 10^{-12} watt of 1,000-cycle power, which is 90 db below 1 milliwatt.

The results of this survey indicated that in consideration of the layout and levels of the existing message circuits and of the noise existent on these circuits and on the program circuits, the value for maximum program volume, should, under normal conditions, be + 3 referred to reference volume; that is, at this value the best balance between program to message crosstalk and program circuit noise would result. It was also determined that on very long sections, or on sections where all circuits were subjected to severe noise exposure, the maximum volume on the program circuits could be increased 3 db to improve the signal-to-noise ratio on the program circuits. This higher volume could be permitted in these cases since on the longer sections the message circuits also usually operate at higher levels, and on the especially noisy short sections the increased crosstalk to the message circuits will ordinarily be masked by the greater noise.

The average noise measured at San Francisco or Chicago at the circuit terminals at the reference volume point was 49 db above reference noise "program weighting" when the restoring network was included at the receiving terminal. The noise averaged 5 db higher than this with the restoring network removed. This value of noise is about 43 db below the maximum power of the program measured at the same point with the same measuring instrument. This, therefore, establishes a signal-to-noise ratio of about 43 db, thus permitting a volume range of approximately 40 db.

The various tests referred to gave statistical data concerning the transmission performance of the circuits from which it could readily be predicted that the circuits would transmit programs with very little impairment to quality. To substantiate this, very critical listening tests were made, comparing the quality of a program after it had been transmitted over various length circuits with the same program transmitted over a reference circuit which was distortionless over the frequency range for which the circuits were designed, namely, to 8,000 cycles. Figure 25 shows schematically the terminal arrangements employed at San Francisco for these listening, or, as they are more commonly called, comparison tests.

Various types of programs were used, such as speech, vocal and

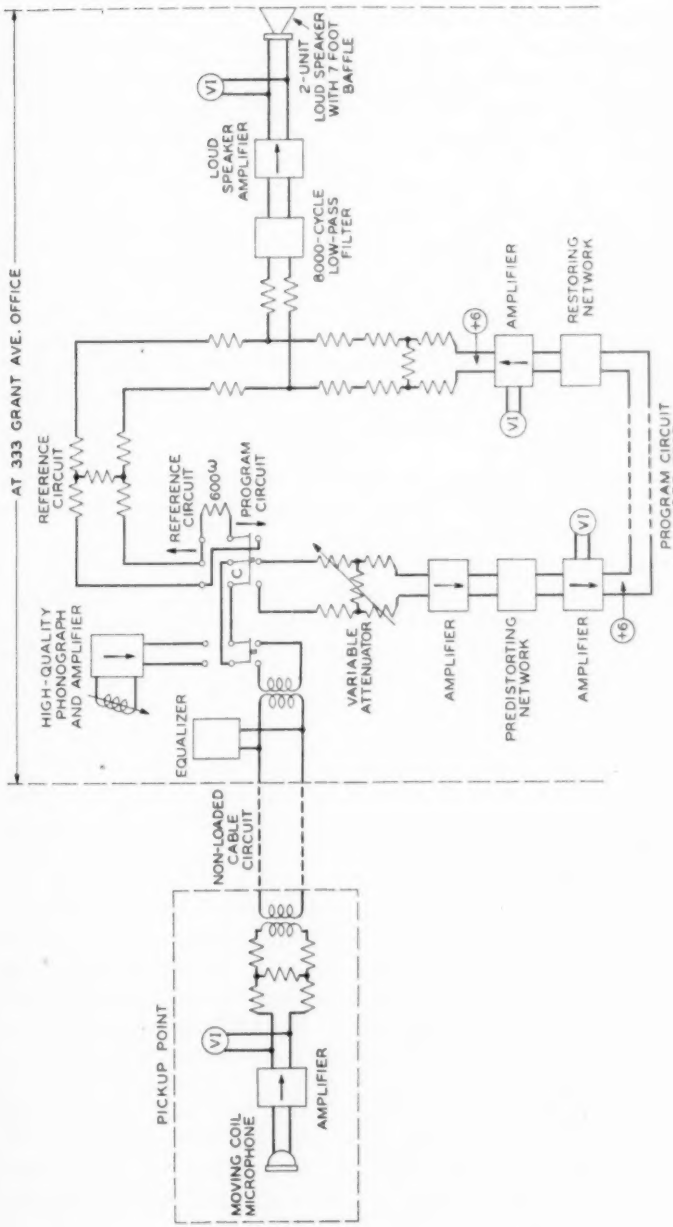


Fig. 25—Terminal arrangements at San Francisco for comparison tests.

instrumental selections, orchestral renditions, both classical and jazz. Quite a number of observers were employed, some of whom were present on several tests and a few on all tests. On tests made on a San Francisco–Denver–San Francisco loop involving 2,600 miles of circuit, no observer was able to consistently differentiate between the quality over the reference circuit and that over the program circuit. On tests made on the San Francisco–Chicago–San Francisco loop certain of the more experienced observers were able to differentiate between the circuits somewhat more than 50 per cent of the time, but this, it must be remembered, was on a direct comparison test. None of the observers could tell with any assurance which was the program circuit and which was the reference circuit if a few minutes were allowed to elapse between switches. On the Chicago loop 264 observations were made on direct comparison tests on which 60 per cent of the observations favored the reference circuit and 40 per cent favored the program circuit.

Included as part of the overall program, were tests to determine the volume range, maximum volume obtainable and speed with which the circuits could be reversed.

On the volume range tests a source of program was obtained and so regulated that it had a very narrow volume range. This was then applied to the circuit with the sending end gain adjusted so that the maximum volume applied at the repeater outputs was + 6. The sending end gain was then gradually decreased so as to apply a gradually decreasing volume to the circuit. This process was continued until the program volume was so weak that the line noise interfered with its satisfactory reception. The amount that the sending end gain was adjusted determined the volume range. The average value for several tests was slightly in excess of 40 db. The maximum volume was determined by switching a 10 db pad from the sending end to the receiving end of the circuit and listening to a transmitted program, noting the point at which there was a quality difference between the high volume and low volume condition. It was found that a slight difference could be detected when the maximum volume on the high volume condition was + 10, thus showing the circuit was capable of handling a maximum volume slightly lower than this value.

As mentioned earlier, switching means are provided at each station for reversing the direction of transmission. On the initial field tests it was demonstrated that the circuits could be reversed readily and at the same time maintain satisfactory overall characteristics. At the present time, on receipt of proper advance notice, the circuits are being reversed on commercial programs in approximately 30 seconds.

Conclusion

The above development provides a program transmission system applicable to open-wire lines which even for very long distances will provide transmission characteristics which should be adequate for program transmission for a number of years to come.

Acknowledgment

The author makes grateful acknowledgment to his associates in the Bell Telephone Laboratories, to members of the Long Lines Department of the American Telephone and Telegraph Company, and of the Pacific Telephone and Telegraph Company, for their cooperation in connection with the setting up of the circuits and participation in many of the tests, and to the National Broadcasting Company and the Columbia Broadcasting System for their assistance in making available the program pickup sources at San Francisco.

Line Filter for Program System *

By A. W. CLEMENT

Open wire circuits recently have been developed for transmitting radio broadcast programs with greater naturalness and over greater distances than heretofore.[1] The simultaneous utilization of these circuits for the transmission of broadcast programs and carrier telephone messages requires the use of line filters to restrict the program and carrier currents to the proper circuits. The low pass line filter developed for the program circuits and its contribution to the maintenance of good quality in the programs transmitted are described in this paper.

PROGRAM transmission systems operated on open wire telephone lines ordinarily are not assigned the exclusive use of the lines, but usually share them with other communication facilities. The wide-band system described in an accompanying paper [1] transmits currents in the frequency band extending from 35 to 8,000 cycles per second, while the lines over which it is routed possess useful transmission ranges extending from 35 to considerably above 30,000 cycles. In order that the range above 8,000 cycles shall not be wasted, carrier telephone systems utilizing these frequencies usually are operated on the same wires with the program systems.

Line filters are used at each terminal and repeater point in the program system to separate the program currents from the carrier currents and to guide each to the proper channel. They are operated in sets consisting of a low-pass filter and a high-pass filter connected in parallel at one end, the end that faces the line. The low-pass filter transmits the program currents freely while effectively excluding the carrier currents, and the high-pass filter transmits the carrier currents while excluding the program currents.[2]

The line filters are located in the open-wire program systems as shown in Figs. 1, 15, 16, 17, 18, and 20 of the accompanying paper by H. S. Hamilton.[1] The low-pass filter is in the direct path of the program currents and therefore has a number of features of special interest. It is the object of this paper to describe this filter and its contribution to the maintenance of good quality in the programs transmitted over the system.

* Published in April, 1934 issue of *Electrical Engineering*. Scheduled for presentation at Pacific Coast Convention of A. I. E. E., Salt Lake City, Utah, September, 1934.

[1] "Wide-Band Open-Wire Program System" by H. S. Hamilton, published in this issue of the *Bell. Sys. Tech. Jour.*

[2] "Telephone Transmission Networks" by T. E. Shea and C. E. Lane, published in *A. I. E. E. Transactions*, Vol. 48, 1929, pages 1031–1034.

This low-pass line filter, with its associated high-pass filter, makes it possible to use the open-wire lines simultaneously for wide-band program service and for commercial carrier telephone service, without impairing the quality of the program. It represents an improvement over older types of line filters, as well as an advance in the technique of equalization in filters. In cases requiring careful delay and loss equalization, it has been the usual practice to design the filter first to supply the required discrimination or filtering action, and then design a delay corrector to correct for the delay distortion in the filter, after which a loss equalizer is designed to correct for the amplitude distortion in both the filter and the delay corrector. The loss equalizer introduces a small delay distortion which usually can be anticipated and corrected in the delay corrector. In the wide-band program filter the functions usually performed by these three separate types of networks have been combined, with a consequent saving in cost and space.

Requirements to Be Met by Program Filter

To function effectively as a line filter, the low-pass filter must provide sufficient discrimination against carrier currents to make their effect completely inaudible in all the receivers connected to the program system. Discrimination varying from 46 to about 90 db is necessary to accomplish this end. Because of the presence of an auxiliary low-pass filter [1] which supplies considerable loss in the frequency ranges where the requirement is unusually severe, each line filter need furnish discrimination varying only from 40 to about 60 db.

From the standpoint of program quality, it is essential that the line filter, while furnishing the foregoing discrimination, shall not introduce any appreciable distortion into the program. This requirement would call for nothing unusual in the way of filter design if there were only a few filters in the system. Long open-wire program systems, however, may extend as far as 3,000 or 4,000 miles, and may contain as many as 50 low-pass line filters. A program that has traversed such a circuit still must be comparable in quality to a program that is broadcast from the point at which it originated. Since the system contains much other apparatus, such as equalizers and amplifiers, each low-pass line filter can be permitted to introduce not more than about 1/100 of the distortion that can be tolerated in the whole system, assuming 50 filters in the system.

There are two types of distortion that must be controlled very carefully in the program filter: these are (1) amplitude distortion, and (2) delay, or phase, distortion. Amplitude distortion is introduced

by a filter when its loss is not the same at all frequencies in the transmitted band, currents of some frequencies being attenuated more than others. The effect of amplitude distortion on the program is to change the relative intensities, or volumes, of tones of the frequencies at which distortion occurs, thus impairing the naturalness of the program. Amplitude distortion ordinarily can be corrected without much difficulty by means of suitable attenuation equalizers.

Delay distortion is introduced by a filter when different frequency components of a signal require different lengths of time for propagation through the filter. This type of distortion is related directly to the shape of the phase shift-frequency characteristic. The slope of this phase shift curve usually is taken as a measure of the delay introduced by the filter. Stated mathematically, the delay in seconds is taken as $\partial B/\partial\omega$, where B is the phase shift in radians and ω is $2\pi f$, f being the frequency in cycles per second. Thus if the phase shift of the filter is proportional to frequency, $\partial B/\partial\omega$, or the delay, is constant and there is no delay distortion. In this case the wave form of a signal transmitted through the filter remains unchanged, the signal being delayed in transmission an interval of time corresponding to the slope of the phase shift curve. If the slope of this curve is not constant over the transmitting band of the filter, however, delay distortion is introduced. In low-pass filters, the difference between the slope of the phase shift curve at a given frequency and the minimum slope of the curve is a measure of the delay distortion at that frequency.

A discussion of delay distortion in telephone apparatus, including filters, as well as a discussion of the effect of delay distortion on telephone quality, may be found in two recent articles on these subjects.[3,4] Whereas the effect of amplitude distortion is to weaken or strengthen some of the tones in the sound being transmitted with respect to the other component tones, the effect of delay distortion is to introduce unnatural audible effects which may become so pronounced as to be annoying if the delay distortion be great enough.

Delay distortion is present in most filters used in communication work, but ordinarily not in such magnitude that its effect is noticeable. As a rule, it need be considered only when a large number of filters is used in a single circuit, as in the case of the program systems. Delay distortion is in general more difficult to correct than amplitude distortion. One of the unusual features of the low-pass line filter used in the wide-band program circuits is the means employed to keep it free from delay distortion.

[3] "Phase Distortion in Telephone Apparatus" by C. E. Lane, *Bell. Sys. Tech. Jour.*, July, 1930.

[4] "Effects of Phase Distortion on Telephone Quality," by J. C. Steinberg, *Bell Sys. Tech. Jour.*, July, 1930.

The filter consists of four parts, each with distinguishing functional characteristics. The separate parts, or sections, have image impedances such that when they are joined together no current is reflected at the junctions. Figure 1 shows the filter in block sche-

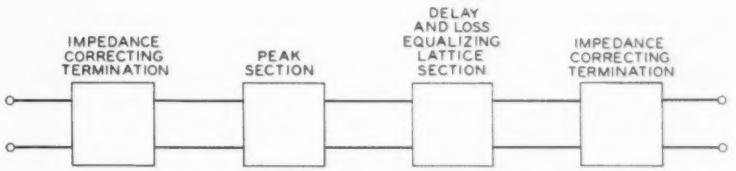


Fig. 1—Block schematic diagram of filter.

matic form. Each part of the filter provides some of the attenuation required to exclude carrier currents from the program circuit, the attenuation of the complete filter being the sum of the attenuations of all parts. On Fig. 2 are shown the loss-frequency characteristics of the various sections and of the complete filter.

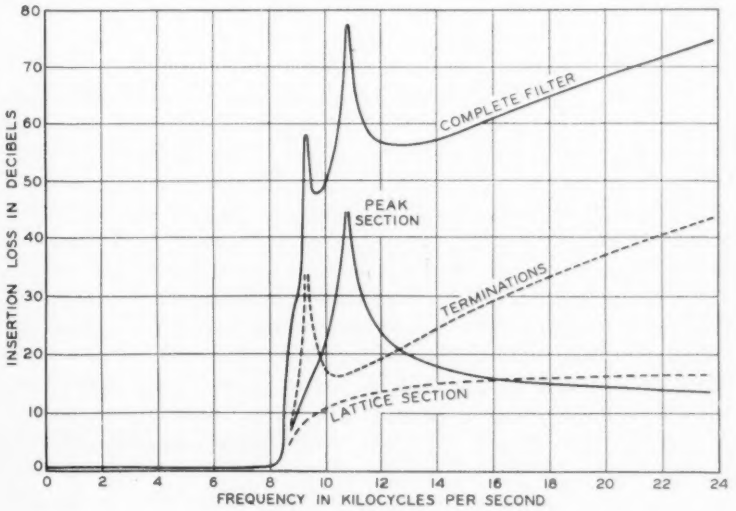


Fig. 2—Loss in filter and in component sections.

Delay Equalization

Likewise, the phase shift of the complete filter is the algebraic sum of the phase shifts of all sections. The phase shift of the filter exclusive of the delay and loss equalizing section is similar to that of the

usual ladder type low-pass filter. Over the lower frequencies of the transmitting band the phase shift-frequency characteristic is practically linear with frequency, but at the higher frequencies the slope of this curve increases gradually with frequency and becomes very large near the upper edge of the band. Phase shift varying in this manner introduces much more delay distortion than can be tolerated, and therefore has to be corrected. It is one of the functions of the delay and loss equalizing section, which is of the lattice type, to correct for this distortion. The phase shift of this lattice section is such that when it is added to that of the rest of the filter the total phase shift is very nearly proportional to frequency over the whole program band, and delay distortion thus is almost entirely eliminated.

The property of the lattice section by which its phase shift can be made to vary with frequency in the desired manner is expressed in the following characteristic equation, which holds only in the transmitting band and when the section is terminated in its image impedances:[5]

$$\tan \frac{B}{2} = \frac{Kf\left(1 - \frac{f^2}{f_2^2}\right)\sqrt{1 - \frac{f^2}{f_c^2}}}{\left(1 - \frac{f^2}{f_1^2}\right)\left(1 - \frac{f^2}{f_3^2}\right)}. \tag{1}$$

In this equation, B is the phase shift in radians; f is the frequency in cycles per second; f_1, f_2, f_3, and f_c are frequencies at which the phase shift of the section is successive multiples of π radians or 180 deg., f_c being also the cut-off frequency of the filter; and K is a constant controllable by assigning the proper values to the coils and condensers of the section. By assigning to f_1, f_2, and f_3 the values of frequency at which it is desired that the phase shift of the section shall be π, 2π, and 3π radians, respectively, and by giving K the proper value, the phase shift-frequency curve is made to approximate the ideal one which completely would correct the delay distortion of the filter. Figure 3 illustrates the building up of the phase shift characteristic.

The delay corresponding to the rate of change of the phase shift with frequency is plotted in Fig. 4. The average delay introduced by the filter is about 0.00035 sec. It may be noted that for frequencies below 7,500 cycles per second, the variation from this average does not exceed 0.000025 sec. Thus the delay due to 50 filters in a long program circuit does not deviate from the average in this frequency range by more than 0.00125 sec. Distortion of this amount ordinarily would not be detected by the average listener. Above 7,500 cycles

[5] U. S. Patent No. 1,828,454 to H. W. Bode.

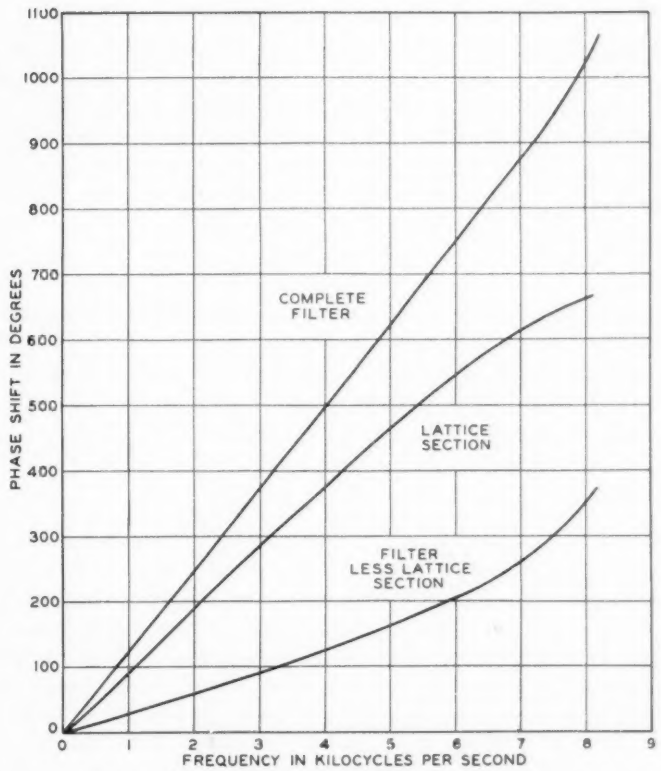


Fig. 3—Phase shift in filter and in component parts.

per second the delay gradually increases with frequency, rising quite rapidly outside the program band. The high attenuation at frequencies above the program range, however, eliminates any effect this distortion otherwise might have on the program.

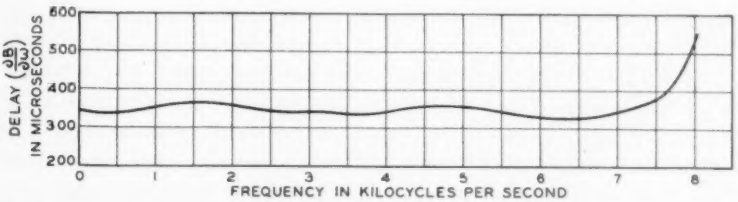


Fig. 4—Delay-frequency characteristic of filter. The ordinates of this curve are proportional to the slope of the upper curve of Fig. 3.

Loss Equalization

Another function of the lattice section is to make the loss of the filter constant in the program frequency band. In a dissipationless filter terminated in its image impedances (which is substantially the condition under which this filter is operated) the loss in the transmitting band is zero. The effect of dissipation is to introduce a loss which is given approximately in this band by the equation:

$$A_d = \frac{\omega}{2Q} \frac{\partial B}{\partial \omega} \tag{2}$$

where A_d is the loss due to dissipation, B is the phase shift of the non-dissipative filter, and Q is the average dissipation factor of the coils (dissipation in the condensers being negligible, ordinarily). The factor Q is equal to the average of the ratios $\omega L_e/R_e$, and $\omega/2Q$ in equation (2) therefore may be written $R_e/2L_e$, where R_e and L_e are the effective resistance and effective inductance, respectively, of the coils.

In the coils of the program filter, Q is about proportional to frequency over the lower portion of the program band, but above this range the factor $\omega/2Q$ increases with frequency. For the filter exclusive of the lattice section, the factor $\partial B/\partial \omega$ is also greatest at the higher frequencies, as may be seen from the lower curve in Fig. 3; hence this part of the filter introduces much more amplitude distortion than is permissible. For the lattice section alone, however, the factor $\partial B/\partial \omega$ is greatest at the lower frequencies, as is apparent from the middle curve of Fig. 3. Thus the natural tendency of dissipation in the lattice section is to compensate for the distortion in the other sections of the filter. This compensating tendency can be controlled to a considerable degree, since by equation (2) A_d is proportional to R_e. By proper adjustment of the effective resistance of the coils of the lattice section, its loss is made practically complementary to that of the rest of the filter, so that the loss of the complete filter is substantially constant throughout the program range.

The loss of the filter in the transmitting frequency band is shown in Fig. 5. The average loss below 7,000 cycles per second is about 0.53 db and the deviation from this average does not exceed 0.03 db. Considering again a circuit containing 50 filters, the deviation from the average loss introduced by the filters does not exceed 1.5 db in this range. Between 7,000 and 7,500 cycles per second the amplitude distortion per filter is about 0.10 db, and above 7,500 cycles the loss increases in such a way as to tend to mask the small delay distortion in this range.

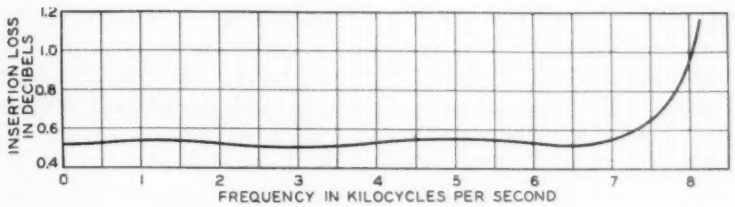


Fig. 5—Loss of filter in program frequency band.

Impedance Correction

In the discussion of the lattice section it was stated that its phase shift is given by equation (1) only when the section is terminated in its image impedance. To facilitate the design and simplify the filter structure, this section has been given an image impedance of the simplest type. This impedance, Z_I, varies with frequency according to the following equation:

$$Z_I = \frac{Z_o}{\sqrt{1 - \frac{f^2}{f_c^2}}}, \tag{3}$$

where Z_o is the "nominal impedance" of the filter, a constant equal approximately to the average impedance of the open-wire lines in the program band; and f_c is the theoretical cut-off frequency. Thus the image impedance rises with increasing frequency to a very high value near the cut-off; and, since the line impedance is practically constant except at very low frequencies, a large mismatch would result at the upper edge of the transmitted band if the lattice section were connected directly to the line. The impedance correcting sections at the ends of the filter are employed to avoid this mismatch. The properties of these sections are such that when they are inserted between the lattice section and the line or the office terminating apparatus, the impedance of the filter matches that of the line and the office apparatus, and the lattice section faces its own image impedance. In this manner, both internal and external reflections largely are avoided; and the phase shift of the lattice section has the proper value.[6]

The general theory on which the design of the impedance correcting sections is based is discussed at length in a recently published article.[7] In brief, the sections consist of two parts: a 4-terminal network to

[6] "Impedance Correction of Wave Filters," by E. B. Payne, *Bell. Sys. Tech. Jour.*, October, 1930.

[7] "A Method of Impedance Correction," by H. W. Bode, *Bell. Sys. Tech. Jour.*, October, 1930.

make the resistance of the filter approximately constant over the program band, and a 2-terminal network placed in shunt at the end to cancel the reactance of the filter in this band. The inductance and capacitance of the coils and condensers of the 4-terminal network are related to the coefficients of a power series expansion of the right-hand part of equation (3) in the manner explained in the article by H. W. Bode.[7] The 2-terminal shunt network at the apparatus end is designed so that, while canceling the reactance of the filter in the program band, it resonates just above the band to produce a peak or sharp maximum of attenuation. It thus supplies the sharp selectivity required to produce an abrupt change from free transmission of the program frequencies to high attenuation of the carrier frequencies.

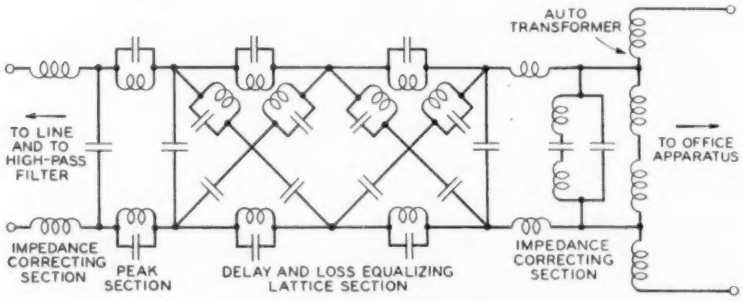


Fig. 6—Schematic diagram of filter.

At the line end, the impedance correcting section is designed for parallel connection with the high-pass line filter. The high-pass filter itself acts as the shunt reactance-canceling network.

The peak section shown at the left of the delay and loss equalizing section in Fig. 1 provides attenuation which rises rapidly with frequency above the program band in such a way as to add to the selectivity of the filter. It is a ladder section of a type often employed in filters for its selectivity.

The filter is designed to match the average impedance of the open-wire lines. The impedance of the office apparatus, however, is slightly higher than that of the lines and the filter. An autotransformer therefore is used at the end of the filter connected to the office apparatus, to effect the required change in impedance. A schematic diagram of the complete filter is shown on Fig. 6, the parts being marked for identification in accordance with the foregoing discussion.

Contemporary Advances in Physics, XXVIII
The Nucleus, Third Part *

By KARL K. DARROW

Transmutation, the major subject of the Second Part of this sequence on the nucleus, assumes again a leading role in the present article. Remarkable cases have been discovered since the first of the year, including a great number in which the impact of one nucleus upon another (or of a neutron on a nucleus) provokes an instantaneous transmutation which is followed after seconds, minutes or hours by the spontaneous breaking-apart of one of the resultant nuclei. One may say that these last are the nuclei of new kinds of radioactive elements, and the phenomena are often called "induced radioactivity"; but many of these new unstable elements differ from all radioactive bodies hitherto known in that they emit *positive* electrons. Some additional examples of transmutation are described at the end of this article.

Induced Radioactivity

UP to the end of last year (1933) it was taken for granted that transmutation is practically instantaneous: that when two nuclei collide, the ensuing fusion and disruption (if any there be) are ended within a time inappreciably short. Nowadays, however, many cases are being discovered, in which a disruption occurs a long time—several minutes or even hours, possibly not for days—after the collision. We must suppose that at the moment of the collision something happens, which entails the eventual disruption. In a very few cases we may be reasonably sure that this initial "something" is itself a transmutation, resembling those previously known in that it is instantaneous, but differing from them in that one of the resulting fragments is an unstable nucleus, of which the eventual spontaneous disruption is that which is observed. This may be the course of events in all cases, but it is also conceivable that in the collision one of the original nuclei may be put into an unstable state without the occurrence of an initial transmutation.

The first-to-be-known of these phenomena was discovered by M. and Mme. Joliot at the very start of 1934, when they exposed samples of aluminium (and boron and magnesium) to the bombardment of the 5.3-MEV alpha-particles from polonium, and after a few minutes of exposure removed them from the bombarding beam and placed them

* In this issue is published the first section of "The Nucleus, Third Part." The paper will be concluded in the October, 1934 issue.

"The Nucleus, First Part" was published in the July, 1933 issue of the *Bell Sys. Tech. Jour.* (**12**, pp. 288–330), and "The Nucleus, Second Part" in the January, 1934 issue (**13**, pp. 102–158).

beside a Geiger-Müller counter.[1] Hundreds of counts per minute disclosed the emergence of fast-flying particles from the samples. The number per minute fell off *exponentially* (Fig. 1) with the lapse of time: a very important feature, for this is the law of radioactivity. The exponential decline implies that the nuclei which were destined to emit these particles were formed at the moments of collisions and existed intact for periods of time—"lifetimes"—not the same for all but distributed in a perfectly random fashion. Such a decline is

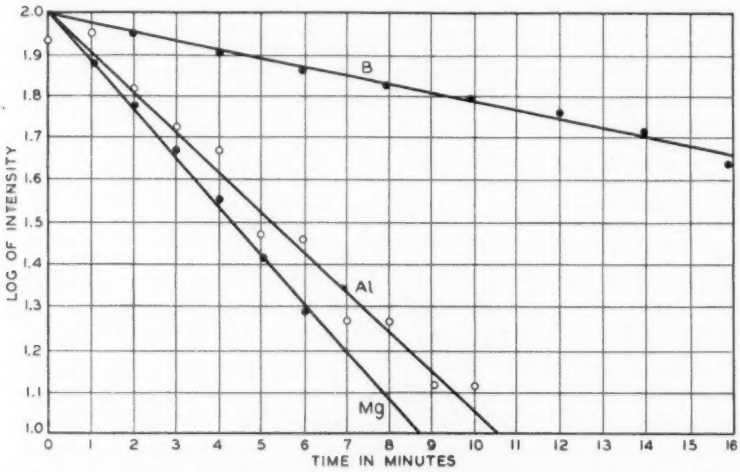


Fig. 1—Exponential decay of the radioactivity induced by boron, aluminium, magnesium with alpha-particles: semi-logarithmic plot. (F. Joliot & I. Curie-Joliot, *Journal de Physique.*)

characterized by a singe constant, the "half-period," or lapse of time during which the rate of emission of particles drops to one-half of its initial value. The half-periods in the three cases examined by the Joliots are different: boron 14′, magnesium 2′ 30″, aluminium 3′ 15″. This is a welcome feature wherever it occurs, as when two substances exhibit different half-periods the effect cannot be ascribed to any contamination common to both.[2]

Since thus there are not only delays between the bombardment and the ultimate disruption, but also (at any rate in the tested cases) a

[1] "The Nucleus, Second Part," p. 119; the "Geiger-Müller" counter has a thin wire for its inner electrode, while most of those called simply "Geiger counters" have needle-points, though the earliest counters invented by Geiger were of the former type.

[2] Cases are on record in which several different elements have exhibited decay-curves, each the sum of two exponentials, one having a half-period characteristic of the element and the other a half-period common to all samples; the latter is then ascribed to a common admixture.

random distribution of the lengths of these delays, it is customary and proper to refer to these phenomena as "induced radioactivity."

Examples of induced radioactivity have already been provoked with all of the four known agents of transmutation: alpha-particles acting on B, Na, Mg, Al and P,—protons acting on boron and carbon—deutons acting on boron and carbon and a number of others—neutrons acting on a large variety of elements. The half-periods reported when neutrons are the agents have ranged from a few seconds to a couple of days, while in all other cases they are of the order of a few minutes.

The nature of the ejected particles resulting from the ultimate disruption is of course of the greatest importance. The Joliots found them to be positive electrons or orestons [3] in their pioneering experiments, and this was confirmed by Ellis and Henderson at the Cavendish Laboratory; the tests have been made by applying magnetic fields to tracks made visible in the Wilson chamber or to beams of particles on their way to photographic or other detectors, and are doubtless to be regarded as conclusive, though no details have yet been published. Induced radioactivity provoked by α-particles, in the few cases so far known, thus results in the emission of orestons.[4] This seems also to be the rule when it is provoked by deutons or protons, as is shown by splendid Wilson-chamber photographs (Figs. 2, 4) obtained by Anderson when samples of various elements (boron in the form of B_2O_3, carbon, aluminium, beryllium) were first bombarded for several minutes and then put right into the chamber itself. The tracks of the particles springing from the samples have the specific aspect of electron-tracks,[5] and in the imposed magnetic field of 800 gauss they have a curvature of which the sense proves the particles to be positive. On the other hand it is stated by Fermi that the radioactivity induced by impacts of neutrons involves the emission of *negative* electrons, though in his very brief reports there is no intimation as to how this is shown.

For each individual case it is important to inquire whether the half-period is independent of such circumstances as the kinetic energy K_0 of the impinging particles. If so, it is sufficient to postulate a single kind of unstable nucleus resulting from the collisions; otherwise, not. This has been investigated in the cases of radioactivity induced by alpha-particle impact; the Joliots reduced K_0 from 5.3 to 1 MEV, without observing any change in the half-perod.

[3] As an occasional alternative to "positive electron" I adopt Dingle's beautiful word "oreston" (Orestes, in Greek mythology, was the brother of Electra).

[4] Excepting that the Joliots have lately reported that magnesium emits electrons of both signs, which they attribute to different isotopes.

[5] "The Nucleus, First Part," p. 303.

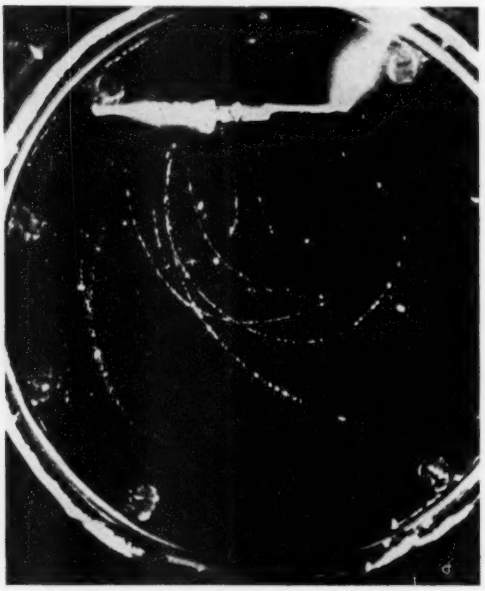


Fig. 2—Induced radioactivity resulting from bombardment of carbon by 0.9-MEV protons: tracks of positive electrons. (C. D. Anderson.)

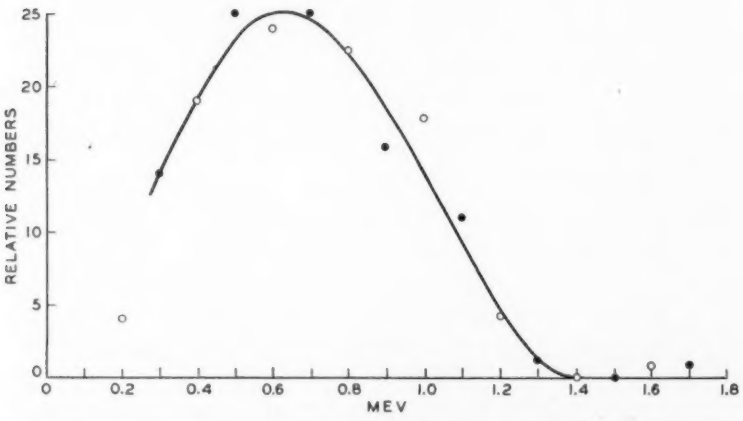


Fig. 3—Distribution-in-energy of positive electrons of the induced radioactivity resulting from bombardment of carbon by 0.9-MEV protons. (Anderson & Neddermeyer, *Physical Review*.)

One next inquires whether all of the orestons resulting from a given type of impact spring off with the same energy. Experience with natural radioactivity shows that while alpha-particles are emitted either with a single definite energy or with one of several definite discrete energies characteristic of the particular process, negative electrons (beta-particles) are always emitted with a very wide and

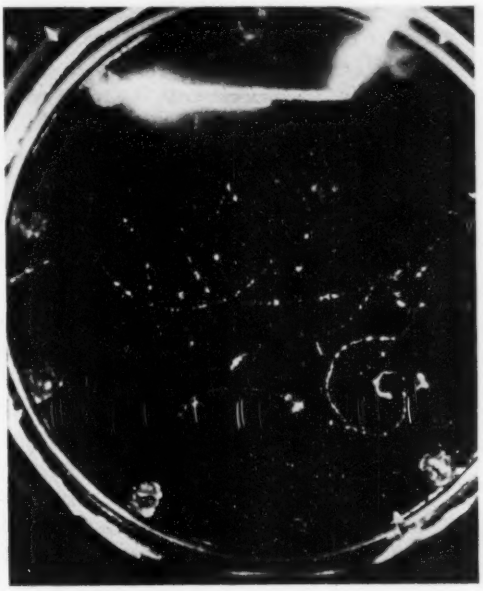


Fig. 4—Induced radioactivity resulting from bombardment of boron oxide by 0.9-MEV deutons: tracks of positive electrons, some springing from gas adjoining the target, as though a radioactive gas had diffused out of the boron oxide block. (Anderson.)

continuous distribution-in-energy. Short as is the time which has elapsed since January last, and weak as are the beams of positive electrons resulting from induced radioactivity, it is already assured that in several cases at least it is the latter rule which is followed and not the former. The best distribution-curves are those derived at Pasadena from a statistical study of oreston-tracks made visible in a Wilson chamber and curved by an imposed magnetic field; they refer to radioactivity provoked by 0.9-MEV protons falling on carbon, and by 0.9-MEV deutons falling on Be, B, C and Al. I reproduce one of these curves as Fig. 3 (another curve obtained with 0.7-MEV protons

falling on carbon is indistinguishable from it). In one of the cases of radioactivity induced by alpha-particle impact, Ellis and Henderson at the Cavendish Laboratory observed a continuous distribution of energies of the positive electrons ranging between 1 and 2.5 MEV.

In all of these cases of delayed transmutation, nothing is observed of the ultimate disruption excepting the emergence of the electron; the other fragments apparently do not receive energy enough to make a track or reach a detector, and our knowledge is thus forcedly incomplete as it is with most other examples of transmutation. In respect to the initial process occurring at the collision, the prospect of attaining complete knowledge seems even dimmer. We are not without some guidance, for when alpha-particles impinge on aluminium or boron, certain particles are expelled with apparently no delay, and these may be fragments resulting from that initial process. There is, however, an *embarras de choix;* both protons and neutrons are expelled in each of these cases; if one is a fragment resulting from the same process of which an unstable nucleus of half-period 3′ 15″ is another fragment, then the other must be due to something entirely different. Actually Ellis and Henderson inferred from their data that in the case of aluminium, the number of protons produced by a given bombardment is fifty times as great as the number of unstable nuclei which eventually eject orestons. This obliges us to assume that the initial process out of which the delayed transmutation arises is either the one which produces the neutrons, or else some other producing no fast-moving particle at all.

Decision between these alternatives is made from a most notable experiment of the Joliots, sufficient indeed by itself to settle the nature of the initial process. To introduce it in the way in which it suggested itself to them, I make the tentative assumption that the initial process is a case of what is called [6] "disintegration by capture with emission of a neutron," and that the residue of this process is the unstable nucleus. Embodying this assumption in equations of "nuclear chemistry" written after the fashion of those in the Second Part with atomic number for a subscript preceding the symbol of each element (so that ${}_0n$ and ${}_1e$ become the proper symbols for a neutron and an oreston) we have for boron and for aluminium:

$$ {}_2\mathrm{He} + {}_5\mathrm{B} = {}_7\mathrm{N} + {}_0n, \quad \textit{followed by} \quad {}_7\mathrm{N} = {}_6\mathrm{C} + {}_1e, $$
$$ {}_2\mathrm{He} + {}_{13}\mathrm{Al} = {}_{15}\mathrm{P} + {}_0n, \quad \textit{followed by} \quad {}_{15}\mathrm{P} = {}_{14}\mathrm{Si} + {}_1e. $$

The unstable nucleus, if it is surrounded by its proper quota of orbital

[6] "The Nucleus, Second Part," pp. 147–148, 155.

electrons, should then possess the chemical properties of nitrogen in the former case, phosphorus in the latter.

The important experiment of the Joliots consisted in showing that when a sample of boron (or aluminium) is first exposed to alpha-particle bombardment and then to such chemical processes as would remove nitrogen (or phosphorus) commingled with the boron (or aluminium), the induced radioactivity is itself removed and carried away. I quote *verbatim:* "We have irradiated the compound BN. By heating boron nitride with caustic soda, gaseous ammonia is produced. The activity separates from the boron and is carried away with the ammonia. This agrees very well with the hypothesis that the radioactive nucleus is in this case an isotope of nitrogen. When irradiated Al is dissolved in HCl, the activity is carried away with the hydrogen in the gaseous state, and can be collected in a tube. The chemical reaction must be the formation of PH_3 or SiH_4. The precipitation of the activity with zirconium phosphate in acid solution seems to indicate that the radio-element is an isotope of phosphorus."

The assumed equations are thus substantiated in a very striking way. These experiments are in a sense the first chemical identifications of any product of transmutation; I say "in a sense," because while this nitrogen and this phosphorus are identified by virtue of chemical properties, they are detected only by virtue of their radioactivity.[7]

Some striking photographs, taken at Pasadena with an expansion-chamber containing a block of boron oxide previously bombarded by alpha-particles, show many tracks of positive electrons springing from points in the air of the chamber (Fig. 4). It is inferred that the unstable nuclei formed from the boron (not from the oxygen, since bombardment of SiO_2 has no effect) are carbon nuclei which unite with electrons to form carbon atoms and then with oxygen atoms to form molecules of CO or CO_2 having a natural tendency to diffuse out of the solid mass. The radioactivity may be driven completely out of the solid block in short order by heating to 200° C. The radioactive particles are unable to pass through a liquid-air trap.

[7] Inserting mass-numbers into the equations, one finds that since Al has but the one known isotope 27, the value 30 is indicated for the mass-number of "radio-phosphorus," as Joliot calls it; while since boron has two isotopes 10 and 11, the two values 13 and 14 are indicated for radio-nitrogen, with no certain evidence to dictate a choice between them. Ordinary stable phosphorus has no known isotope 30, and ordinary stable nitrogen has no known isotope 13, but the vast majority of its atoms are of mass-number 14. It seems natural that a very unstable isotope should have a different mass-number from any of the known and stable ones, and this may be a valid argument for inferring that it is B^{10} rather than B^{11} which is concerned in the induced radioactivity of boron; but there is nothing to prohibit us from supposing that there may be an unstable isotope of nitrogen agreeing in mass-number with the one which is durable.

The result of bombarding carbon with deutons might be expected to be the same as that of bombarding boron with alpha-particles, it being natural to assume the reactions:

$$_1H^2 + {}_6C^{12} = {}_7N^{13} + {}_0n^1, \quad \textit{followed by} \quad {}_7N^{13} = {}_6C^{13} + {}_1e.$$

The half-period of the delayed disruption has been determined at Pasadena as 10.3 minutes. This does not agree with that observed by the Joliots when alpha-particles are projected against boron. The disagreement is not so welcome as agreement would have been, but does not in the least invalidate the foregoing equations, since it is perfectly conceivable that two different unstable nuclei with different half-periods might both have the atomic number 7 and the mass-number 13. Bombardment of carbon with protons leads to delayed disruptions with the same half-period of about ten minutes, and this is not so easy to understand as it may seem, since the obvious notion that the proton and the C^{12} nucleus simply merge into a nucleus N^{13} which later on explodes leads into difficulties with the principles of conservation of energy and conservation of momentum.

As to the way in which the number of observed disruptions varies with the kinetic energy K_0 of the impinging particles, there are data relating to the bombardment of aluminium by alpha-particles. The Joliots varied K_0 from 5.3 MEV downwards; they report that the number of positive electrons diminishes with falling K_0, becoming imperceptible for boron at about 3 MEV, for Mg and Al at 4 to 4.5 MEV. Ellis and Henderson varied K_0 from 5.5 upward to 8.3 MEV, by using alpha-particles emitted from other radioactive bodies than polonium; they found the number of orestons steadily increasing with rising K_0, rising in the ratio 15 : 1 as K_0 was raised from 5.5 to 7 MEV, and showing signs of approaching a maximum not far beyond K_0 = 8.3 MEV.

The positive electrons emitted in induced radioactivity are frequently—perhaps generally—accompanied by high-frequency photons, of which energy-measurements may hereafter show that they are due to the coalescence of positive with negative electrons to form light.

I close this section by listing the elements which have been observed to display induced radioactivity after bombardment by one or other of the four agents of transmutation, and add those which have been tested without positive results, in order to show the scope of the experiments. In certain cases positive results have been obtained by some observers and not by others, but this may signify simply a weaker bombarding stream or a less sensitive detector in the apparatus of the latter.

Bombardment by alpha-particles: B, Mg, Al (Joliots, Ellis & Henderson); Na, P (Frisch); negative results with H, Li, Be, C, N, O, F, Na, Ca, Ni, Ag (Joliots).

Bombardment by deutons: Li, Be, B, N, C, O, F, Na, Mg, Al, Si, P, Cl, Ca (Henderson, Livingston & Lawrence, with 3-MEV deutons); Li, Be, B, C, Mg, Al (Crane & Lauritsen, with 0.9-MEV deutons).

Bombardment by protons: B, C (Crane & Lauritsen); C (Cockcroft, Gilbert & Walton with 0.6-MEV protons); C (?) (Henderson *et al.*, with 1.5-MEV protons). Negative results by Henderson *et al.* with 1.5-MEV protons on all but C among the elements listed above before their names.

Bombardment by neutrons: F, Na, Mg, Al, Si, P, Cl, Ti, V, Cr, Fe, Cu, Zn, As, Se, Br, Zr, Ag, Sb, Te, I, Ba, La, U (Fermi); F, Mg, Al (Dunning and Pegram).

Other Cases of Transmutation

It is not altogether safe to separate cases of "induced radioactivity" from "other cases of transmutation," inasmuch as most of the latter class have been observed under conditions where it was impossible to tell whether or not there was a delay between collision and disruption, and perhaps some of them belong in the former class. Of certain transmutations one may say that if there is such a delay, the law of conservation of momentum must be suspended for the duration thereof, resuming its sway only at the moment of the disruption. Nevertheless I should not wish to affirm that for the processes mentioned in this section or in the Second Part the delay is always literally zero.

Early in this year was first achieved, at the Cavendish Laboratory by Oliphant, Shire and Crowther, what had been the aim of many physicists for over a decade: the separation of a metal, normally consisting of more than a single isotope, into films each comprising atoms of practically a single isotope only, and thick enough for physical experiments. This was performed with lithium, and when protons and alternatively deutons were projected against films of Li^6 and alternatively Li^7, the four resulting sets of observations settled the attributions of the various groups of fragments previously observed when ordinary blocks of lithium had been bombarded. The origin of the two long-range groups of paired alpha-particles described in the Second Part was precisely as had been suspected: they proceed from the interactions:

$${}_1H^1 + {}_3Li^7 = 2{}_2He^4 + (T_1 - T_0); \quad {}_1H^2 + {}_3Li^6 = 2{}_2He^4 + (T_1 - T_0),$$

where $(T_1 - T_0)$ stands for the amount of energy transformed in each

reaction from energy-of-rest-mass to kinetic energy, equal to about 17 MEV in the first case and to about 23 MEV in the second. The continuous distribution of alpha-particles up to range 7.8 cm (Fig. 9 in the Second Part) is due to impacts of deutons against Li^7, and thus may still be attributed to a transmutation in which three nuclei,—a neutron and two alpha-particles—spring from the merger of a deuton with a Li^7 nucleus. Of the other attributions I shall presently speak.

The transmutations arising from the impact of deutons on deuterium are in some ways unique. They are the first to be known in which the two colliding particles are identical, both being H^2 nuclei; one of them appears to be much the most abundant yet observed, in the sense that a given number of bombarding particles produces an unprecedentedly great number of detectable fragments; each of them results in the formation of a nucleus long sought but never certainly detected till 1934.

The better-known of these reactions is described by the equation,

$$ {}_1H^2 + {}_1H^2 = {}_1H^3 + {}_1H^1 + (T_1 - T_0). \qquad (1) $$

It is both somewhat amusing and somewhat annoying to realize that this is not a transmutation at all in the formerly-proper sense of the word, since there is no change of one element into another! the hydrogen isotope of mass-number 2 is changed into hydrogen isotopes of mass-numbers 1 and 3 respectively; it will be desirable to enlarge the scope of the term "transmutation" to cover cases like this one. The H^1 nuclei resulting from this reaction were vividly demonstrated by Tuve and Hafstad when they projected deutons into divers gases in an ionization-chamber—air, carbon dioxide, ordinary hydrogen, and deuterium successively; there were no emerging protons (of range superior to 3.5 cm, the minimum observable) from any of the three first named, but from the last there was the "very large yield" of one proton per several thousand impinging deutons. Another estimate of yield has been supplied from the Cavendish school, by Oliphant Harteck and Rutherford; theirs refers to impacts by deutons of energy 0.1 MEV, a value considerably smaller than those of Tuve's research; they find that the number of protons coming forth from a thick layer of deuterium is of the order of a millionth of the number of such deutons entering the layer. The estimates do not seem incompatible, especially as the Cambridge people find the number of fragments to be mounting very rapidly as the deuton-energy T_0 increases;[8] and they show that any possibility of a slight admixture

[8] The "thick layers" are films of certain compounds of hydrogen in which a large proportion of the usual H^1 atoms have been replaced by H^2 atoms. The curve

of deuterium with any other substance must be very carefully considered and assessed, whenever that other substance is bombarded with a beam containing deutons and it is observed that protons are produced.

The range of the protons due to the foregoing reaction is about 14 cm when T_0 is low—0.1 MEV or thereabouts—and rises with T_0. Translate its minimum value into the corresponding kinetic energy (obtaining about 3 MEV); compute the momentum of the proton—this, save for a minor correction due to the relatively small momentum of the impinging deuton, should be opposite in direction and equal in magnitude to the momentum of the other fragment of the transmutation, the nucleus H^3. Thence compute the kinetic energy of this other fragment, and estimate thence its presumable range; owing to our lack of experience with such particles the estimate may not be very exact; Oliphant, Harteck and Rutherford arrive at the figure 1.74 cm. Now, the protons of 14-cm range of which I have been speaking are not the only fragments to be observed when deutons impinge on deuterium. There are also particles of a much less range; these are equally numerous with the 14-cm protons, and expansion-chamber photographs by Dee have shown that a track of the one variety is likely to be paired with a track of the other, after the fashion of the paired tracks due to the transmutations H + Li = 2He (Figs. 14 and 15, Second Part); and their range of about 1.6 cm. is taken by the Cavendish people as being in substantial agreement with the estimate aforesaid. It is this interlocking of concordant observations which speaks so strongly for the rightness of this description of the reaction, and therefore for the existence of the hitherto-unknown isotope H^3 of hydrogen.

Meanwhile it has been discovered at Princeton that the new isotope can be generated by maintaining a self-sustaining discharge in gaseous deuterium: a way of achieving transmutation several times attempted in past years, but never (so far as I know) with proved success. Out from the discharge tube (where the voltage is 50,000 to 80,000) some of the ionized atoms and molecules shoot through a hole in the cathode into another and very large chamber filled with deuterium in which they disperse themselves, thus having opportunities for transmutation in both this chamber and the tube. A sample of the gas is afterwards

of number-of-fragments *vs.* T_0 shows the peculiar shape common to such curves when obtained with thick layers, which suggests that as T_0 is raised the increase in the number of transmutations is at first partly due to an increase in the probability of transmutation at an impact, but later entirely due to the fact that the faster particles enter farther into the layer and have more opportunities of striking nuclei before their energy is gone than do the slower (The Nucleus, Second Part, p. 141). The theory of such curves has, however, never been worked out.

extracted and is ionized in a separate chamber; the charge-to-mass ratios of its ions are determined by an especial type of deflection-apparatus. Search is made for ions having the charge-to-mass ratio of a singly-ionized molecule of mass about 5, such as could be a molecule H^2H^3. Such ions occur. To discover them, however, is not the same thing as to prove the existence of H^3, since so far as anyone can tell from their charge-to-mass ratios (as measured with the accuracy attainable in these experiments) these ions might have the constitution $H^2H^2H^1$—there being some of the isotope H^1 in the gas. How to make such discriminations is one of the major problems in the analysis of the ions found in gases. In this case it happens to be known that in ordinary hydrogen, the ratio of the number of triatomic to that of diatomic molecular ions is proportional to the density of the gas. Now in these experiments, the ratio of the number of mass-5 ions to the number of mass-4 ions is the sum of two terms, one proportional to the gas-density and the other independent of it. The latter term is taken as the measure of the amount of H^2H^3, therefore of H^3, in the gas. A like study made with deuterium none of which had been exposed to the discharge indicated a very small amount of H^3, about one atom in two hundred thousand of H^2; the discharge enhanced this ratio fortyfold in an hour.

To return to the work at the Cavendish Laboratory: the lesser-known of the two reactions which may occur when deutons meet is probably described by the equation,

$$ {}_1H^2 + {}_1H^2 = {}_2He^3 + {}_0n^1 + (T_1 - T_0) \qquad (2) $$

and is a transmutation in the strictest sense of the word, helium as well as neutrons [9] appearing out of hydrogen. I refer to it as lesser-known, because although the neutrons have been observed the helium nuclei have not been. This lack of evidence withholds a desirable support from the equation, but does not contradict it; for on measuring the momentum of the neutron, equating it to that of the hypothetical He^3 nucleus and estimating the range of the latter, this range turns out to be so small as to make detection difficult. We are not, however, without other evidence for He^3; when protons are projected against lithium, particles of ranges 1.15 cm. and 0.68 cm. appear,[10] and the observations made with monisotopic films show that Li^6 is involved in their origin: if we suppose

$$ {}_1H^1 + {}_3Li^6 = {}_2He^4 + {}_2He^3 + (T_1 - T_0) \qquad (3) $$

[9] Harkins has suggested the name "neuton" for the element of which neutrons are the ultimate particles.

[10] Kirchner has lately observed an 0.9-cm. group.

the equation is supported by the facts that the ranges of the two groups stand to one another in the ratio computed by assuming equality of momenta, that particles of one are found to be paired with particles of the other, and that they ionize about as much as alpha-particles of equal range.

The rest-masses of the two new nuclei are estimated by putting, in equations (1), (2) and (3), the best available values for T_0 (the kinetic energy of the impinging deuton, that of the other H^2 nucleus being negligible) and T_1 (the sum of the kinetic energies of both fragments resulting from the reaction). The results are: for the rest-mass of H^3, 3.0151 from (1); for the rest-mass of He^3, 3.0166 from (3). To derive the latter from (2) is not so precise, the energy of neutrons being harder to evaluate than that of charge-bearing particles; Oliphant, Harteck and Rutherford prefer to say merely that the result is not incompatible with that from (3).[11]

These are the fourth and fifth of the nuclei (counting the neutron as one) in order of increasing mass. The departures of their masses from the adjacent integer are abnormally great for light nuclei, and their packing-fractions (First Part, p. 318) are the greatest yet known excepting that for H^2, and fall neatly by the upper branch of the curve of packing-fraction *vs.* mass-number (Fig. 8 of the First Part). The contrast between the packing-fractions 55 of He^3 and 5 of He^4 is especially striking. The new nuclei are the first isobars to be discovered of mass-number less than 40, and the first pair to be discovered of which the masses are distinguishable.

Cockcroft and Walton have studied at length the fragments emerging from lithium, boron and carbon bombarded by deutons. Lithium supplies a group and boron a group of protons which may result from the transformation of the lighter into the heavier isotope according to the schemes,

$$ {}_1H^2 + {}_3Li^6 = {}_3Li^7 + {}_1H^1 + (T_1 - T_0), \tag{4} $$

$$ {}_1H^2 + {}_5B^{10} = {}_5B^{11} + {}_1H^1 + (T_1 - T_0), \tag{5} $$

but the two members of each equation (in which all the rest-masses are known by deflection-experiments) do not agree very well. Carbon supplies a group and boron two more groups of protons which cannot be made to fit into such a scheme without postulating emission of gamma-rays to achieve the balancing of masses—an emission for which,

[11] These results are computed by assuming that the values of the rest-masses of H^1, H^2, He^4, Li^6, Li^7 and n^1 given by Aston, Bainbridge and Chadwick are exact, and that no additional fragment (such as a gamma-ray photon) of appreciable energy is emitted at the transmutation.

it is true, independent evidence exists in the case of carbon. Boron supplies a group of alpha-particles which may be due to the reaction,

$$ {}_1H^2 + {}_5B^{10} = 3{}_2He^4 \qquad (6) $$

and which comprises the most energetic subatomic particles yet known, those of the cosmic rays excepted (12.3 MEV, range 15 cm.). Blocks of various heavier elements emit both alpha-particles and protons, of which the amounts both relative and absolute vary tremendously with heat-treatment, degassing, and other circumstances, so that evidently they cannot altogether proceed from the element constituting most of the block, and their origins furnish a severe problem for research.

Electrical Wave Filters Employing Quartz Crystals as Elements

By W. P. MASON

This paper discusses the use of piezo-electric crystals as elements in wave filters and shows that very sharp selectivities can be obtained by employing such elements. It is shown that by employing crystals and condensers only, very narrow band filters result. By using coils and transformers in conjuction with crystals and condensers, wide-band-pass and high and low-pass filters can be constructed having very sharp selectivities. The circuit configurations employed are such that the coil dissipation has only the effect of adding a constant loss to the filter characteristic, this loss being independent of the frequency. Experimental curves are given showing the degree of selection possible.

In the appendix, a study is made of the modes of motion of a perpendicularly cut crystal, and it is shown that all the resonances measured can be derived from the elastic constants and the density of the crystal. The effect of one mode of motion on another mode is shown to be governed by the mutual elastic compliances of the crystal. By rotating the angle of cut of the crystal, it is shown that some of the compliances can be made to disappear and a crystal is obtained having practically a single resonant frequency over a wide range of frequencies. Such a crystal is very advantageous for filter purposes.

Introduction

FILTERS for communication systems must pass, without appreciable amplitude distortion, waves with frequencies between certain limits, and must attenuate adequately all waves with somewhat greater or smaller frequencies. To do this efficiently, the change from the filter loss in the transmission region, to that in the attenuation region, must occur in a frequency band which is narrow compared to the useful transmission band. At low frequencies, ordinary electrical coil and condenser filters can perform this separation of frequencies well because the percentage band widths (ratio of band width to the mean frequency of the band) and the percentage separation ranges (ratio of the frequency range required, in order that the filter shall change from its pass region to its attenuated region, to the adjacent limiting frequency of the pass band) are relatively large.

For higher frequency systems, such as radio systems, or high frequency carrier current systems, the band widths remain essentially the same, and hence the percentage band widths become much smaller. Here separation by coil and condenser filters becomes wasteful of frequency space. For these filters, owing to the relatively low reactance-resistance ratio in coils (this ratio is often designated by the letter Q) the insertion loss cannot be made to increase faster than a

certain percentage rate with frequency. Hence an abrupt frequency discrimination cannot be obtained between the passed frequency range and the attenuated frequency range. In present radio systems, double or triple demodulation is often used to supplement the selectivity of filter circuits.

If, however, elements are employed which have large reactance-resistance ratios, filters can be constructed which have small percentage bands and which attenuate in small percentage separation ranges. Such high Q elements are generally obtainable only in mechanically vibrating systems. Of these elements, probably the most easily used is the piezo-electric crystal, for it possesses a natural driving mechanism.

It is the purpose of this paper to describe work which has been done in utilizing these crystals as elements in filters.[1] Since the quartz crystal appears to be the most advantageous piezo-electric crystal, all of the work described in this paper is an application of this type of crystal. The possibilities and limitations are discussed and experimental data are given showing that these filters are realizable in a practical form.

Piezo-electric Crystals and Their Equivalent Electrical Circuits

When an electric force is applied to two plates adjacent to a piezo-electric crystal, a mechanical force is exerted along certain directions which deforms the crystal from its original shape. On the other hand deformations in certain directions in the crystal produce a charge on the electric plates. Hence the crystal is a system in which a mechanical electrical coupling exists between the mechanical and electrical parts of the system.

Quartz crystals, particularly when vibrating along their smallest dimension, as they do for high frequency oscillators, have a large number of resonances which do not differ much in frequency from the principal resonance. While this is no great disadvantage for an oscillator, since an oscillator can pick out the strongest resonance and utilize it only, the large number of resonances is a great disadvantage when using

[1] The development of ideas in the direction of employing crystals as elements of selecting circuits dates back to Cady who in patent—Re. 17,358 issued July 2, 1929, original filed January 28, 1920—showed various types of tuned circuits of which crystals formed a part. Subsequently Espenschied in patent 1,795,204, issued March 3, 1931, filed January 3, 1927—patented broadly the use of crystals as elements of true filter structures. More recently a patent of the writer's—1,921,035 issued August 8, 1933, filed Sept. 30, 1931—describes the use of crystals in lattice structures, and this patent, together with several others pending, forms the basis for the filters discussed in this paper. It is only within the last few years that filter structures including crystal elements have been practically realized.

the crystal to select currents over a wide band of frequencies and to reject currents whose frequencies lie outside this pass region. Hence it is advantageous for filter uses to obtain a crystal which has substantially a single prominent resonance over a wide range of frequencies. Such a vibrating element can usually be obtained only by making the dimension along which the crystal is vibrating, large compared to the other dimensions, and this fact determines the best cut of crystal to use.

Two principal types of orientations have usually been employed in cutting quartz crystals. The first type is the so-called Curie or perpendicular cut in which the crystal is so cut that its major surfaces are parallel to the optical axis and perpendicular to an electrical axis. Such a crystal is shown by Fig. 1. The second type of cut is the parallel

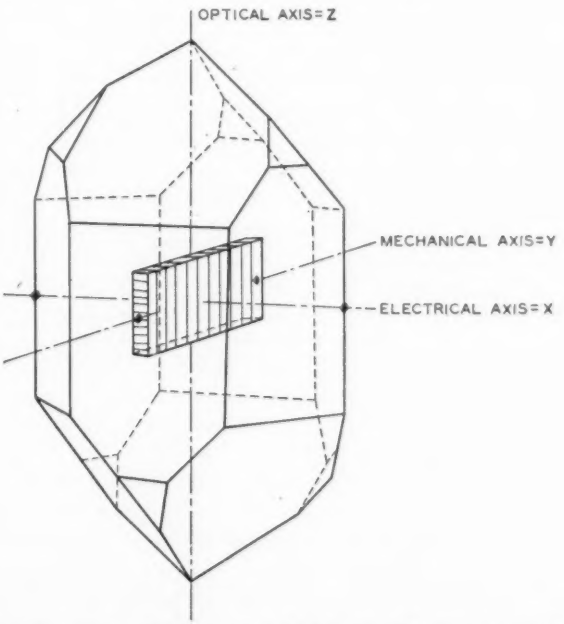


Fig. 1—Orientation of a Curie or perpendicular cut with respect to native crystal.

or 30-degree cut in which the major surfaces of the crystal plate are parallel to both the optical and electrical axes. Since this cut results in a crystal vibrating along its smallest dimension, it is not of much interest for filter uses.

When using a crystal as part of an electrical system, it is desirable

to know the constants of an electrical circuit which has the same impedance characteristic as the crystal. If attention is confined to the single prominent resonance, the electrical circuit representing the crystal is as shown by Fig. 2. Some theoretical consideration has been

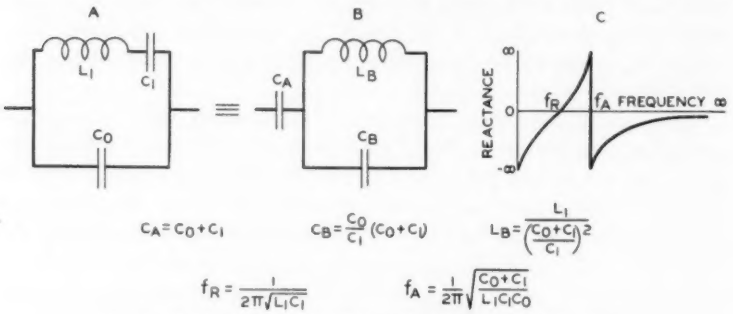


Fig. 2—Equivalent electrical circuit of piezo-electric crystal.

given to the electrical network representing perpendicularly cut crystals by Cady,[2] Van Dyke,[3] Dye,[4] Vigoureux [5] and others. Assuming a quartz plate to have only plane wave motion, Vigoureux has investigated the motion in such a plate, and has shown that there will be resonances at odd harmonics of a particular frequency determined by the length and mechanical constants of the plate. In the neighborhood of the fundamental resonance of the crystal, with the electrical plates placed on the crystal surfaces, he finds the equivalent circuit shown by Fig. 2A, the elements of which in practical units have the following values:

$$
\begin{aligned}
C_0 &= \frac{l_0 l_m K}{4\pi l_e \times 9 \times 10^{11}} = \text{capacitance in farads,} \\
C_1 &= \frac{l_0 l_m 8 E d_{12}^2}{\pi^2 l_e \times 9 \times 10^{11}} = \text{capacitance in farads,} \qquad (1) \\
L_1 &= \frac{l_e l_m \rho \times 9 \times 10^{11}}{8 l_0 E^2 d_{12}^2} = \text{inductance in henries,}
\end{aligned}
$$

where l_0, l_m, l_e are respectively the lengths of the optical, mechanical, and electrical axes in centimeters,

K = specific inductive capacitance = 4.55 for quartz,
E = Young's modulus = 7.85×10^{11} for quartz,
d_{12} = piezo-electric constant = 6.4×10^{-8} for quartz,
ρ = density = 2.654 for quartz.

[2] W. G. Cady: *Phys. Rev.* XIX, p. 1 (1922); *Proc. I. R. E.* X, p. 83, (1922).
[3] K. S. Van Dyke: Abstract 52, *Phys. Rev.*, June, 1925; Proc. I. R. E., June, 1928.
[4] D. W. Dye: *Proc. Phys. Soc.*, XXXVIII (5), pp. 399–453.
[5] P. Vigoureux: *Phil. Mag.*, Dec., 1928, pp. 1140–53.

Inserting these values, the element values in terms of the dimensions become

$$\begin{aligned} C_0 &= 0.402 \times 10^{-12}\, l_m l_0 / l_e, \\ C_1 &= 0.289 \times 10^{-14}\, l_m l_0 / l_e, \\ L_1 &= 118.2\, l_m l_e / l_0. \end{aligned} \tag{2}$$

From these values it is seen that there is a fixed ratio between these two capacitances [6] of

$$r = C_0/C_1 = 140. \tag{3}$$

As will be evident later, this ratio limits the possibilities of the use of quartz crystals in filter circuits.

Experiments with quartz crystals, with electrodes contiguous to the crystal surfaces and with the optical and electrical axes small compared to the mechanical axis, show that these values are approximately correct. The value of C_0 checks the above theoretical value quite closely. The value of C_1 obtained by experiment is somewhat larger than that given by equation (2) and the value of the inductance somewhat smaller. The ratio of C_0/C_1 has been found as low as 115 to 1, but a value of 125 to 1 is about all that can be realized, when account is taken of the distributed capacitance of the holder, connecting wires, etc.

When either of the dimensions along the electrical or optical axes becomes more than a small fraction of the dimension of the mechanical axis, the plane wave equations given above no longer hold accurately. This is due to the fact that a coupling exists between the motion along the mechanical axis and other modes of motion. For an isotropic body, one is familiar with the fact that when a bar is compressed or stretched it tends to stretch or compress in directions perpendicular to the principal direction of motion. This state of affairs may be described by saying that the modes of motion are coupled. In a crystal this same relation exists and in addition, due to its crystalline form, a shearing motion is set up whose shearing plane is determined by the mechanical and optical axes and whose motion is parallel to the mechanical axis. In fact the shearing motion is more closely coupled to the mechanical axis motion that is the extensional motion. As long as the optical axis is small compared to the mechanical axis, this coupling action manifests itself as a decreased stiffness along the mechanical axis, but if a condition of resonance is approached for the motion along the optical axis, the mode of motion is entirely changed. This effect is

[6] In a paper contributed recently to the *Institute of Radio Engineers*, it is shown that this ratio limitation is a consequence of a fixed electro-mechanical coupling between the electrical and mechanical systems of the crystal.

analyzed in the appendix and is quantitatively explained in terms of the elastic constants of the crystal. On the basis of this explanation, an investigation is also given in the appendix, of crystals cut at different orientations, and a crystal having many advantages for filter uses is derived.

Some experimental data [7] have been taken for perpendicularly cut crystals for various ratios of axes. Figure 3 shows the principal reso-

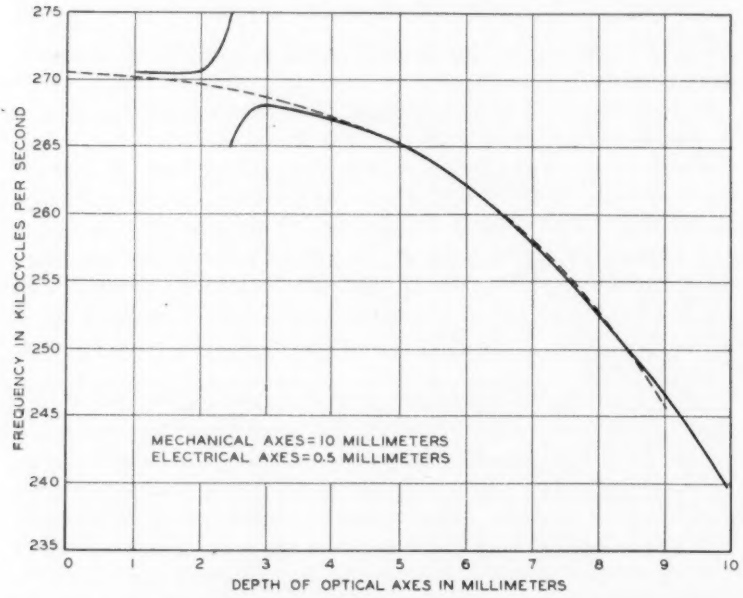


Fig. 3—Principal resonant frequency of a perpendicularly cut crystal as a function of the width of the crystal.

nant frequency (the frequency for which the electrical impedance is a minimum) for a series of crystals whose mechanical axes are all 10 millimeters, whose electrical axes are 0.5 millimeter, and whose optical axes vary from 1 to 10 millimeters. As will be observed, increasing the length of the optical axis in general lowers the resonant frequency. The discontinuity in the curve for the ratio $l_0/l_m = .23$ is discussed in detail in the appendix.

[7] The experimental data shown by Figs. 3 and 4 have been taken by Mr. C. A. Bieling while the temperature coefficient curve of Fig. 5 was measured by Mr. S. C. Hight.

The solid curve of Fig. 4 shows a measurement of the ratio, r, of the capacitances in the simple representation of the crystal shown by Fig. 2A. This ratio is measured by determining the resonant and anti-

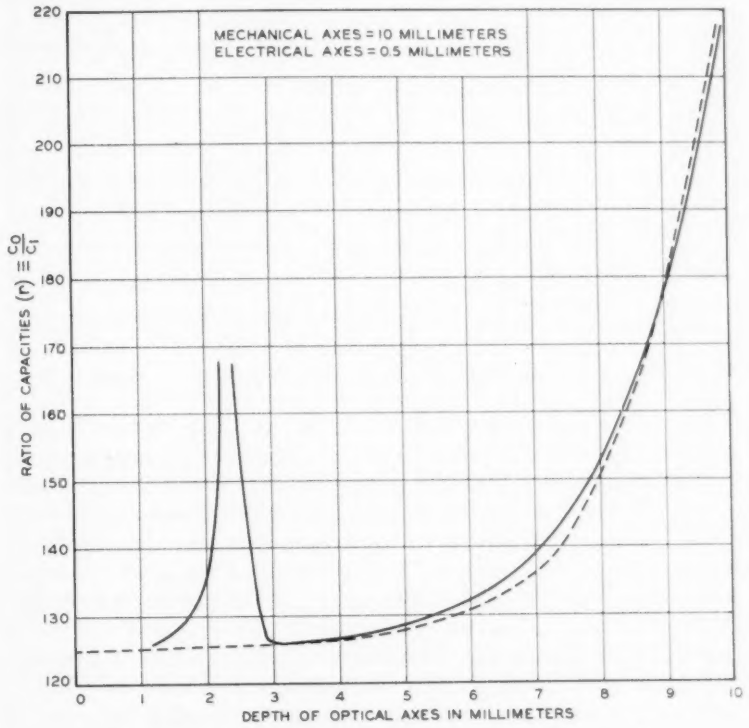


Fig. 4—Ratio of capacitances of a perpendicularly cut crystal.

resonant frequencies of the crystal. r is related to these by the formula

$$f_A^2/f_R^2 = 1 + 1/r, \tag{4}$$

where f_A is the anti-resonant frequency and f_R the resonant frequency.

Figure 5 shows a measurement of the temperature coefficient of the resonant frequency for the same set of crystals. It will be noted that as the optical axis increases in depth, the temperature coefficient increases and that crystals with smaller dimensions along the optical axis in general have much smaller coefficients. Increasing the thickness along the electrical axis has the effect of decreasing the tempera-

ture coefficient and in fact for certain ratios of the three axes the coefficient approaches zero.

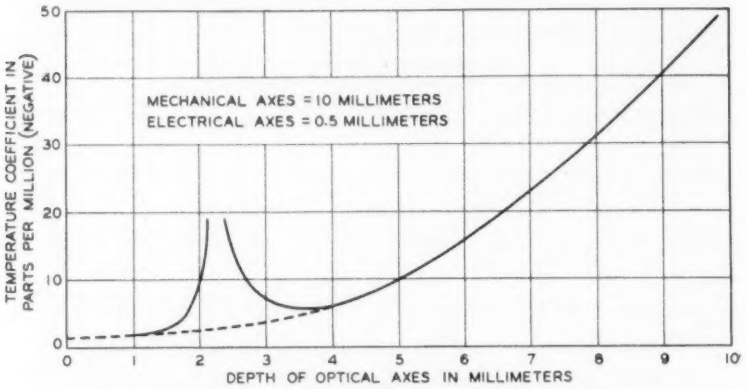


Fig. 5—Temperature coefficient of a perpendicularly cut crystal.

When the crystals are used in filters, two quantities are usually specified, the resonant frequency of the crystal and the capacitance of the series condenser. The shunt capacitance of the crystal is usually incorporated with an electrical capacitance which is specified by other considerations. The resonant frequency is determined principally by the mechanical axis length. The capacitance of the series condenser is determined by the ratio of the area to the thickness or by $l_0 l_m / l_e$. The third condition is given by the fact that the length of the optical axis should be kept as small as possible in order that any subsidiary resonances shall be as far from the principle resonance as possible. The curves of Figs. 3 and 4 and the fact that the resonant frequency of a given shaped crystal varies inversely as the length, can be used to determine the dimensions of the crystal. It is obvious that the crystal should not be used in the region $0.2 < l_0/l_m < 0.3$ on account of the two prominent resonant frequencies.

Use of Crystals and Condensers as Filter Elements

Considering crystals as representable by the simple electrical circuit shown on Fig. 2*A*, these circuits can be utilized as elements in electrical networks. They may, of course, be used in a network employing any kinds of electrical elements. Since, however, their Q is high, it would be advantageous not to employ any electrical elements which do not also have a high Q, in order that the dissipation introduced by these elements may not be a matter of consideration. The

Q's of the best electrical condensers may be as high as 10,000, which is of the same order as the crystal Q, and hence such elements can be employed advantageously with crystals. It is the purpose of this section to discuss the possibilities and limitations of filter sections employing crystals and condensers only.

The simplest types of filter sections are the ladder type networks illustrated by Fig. 6. If crystals and condensers only are employed in

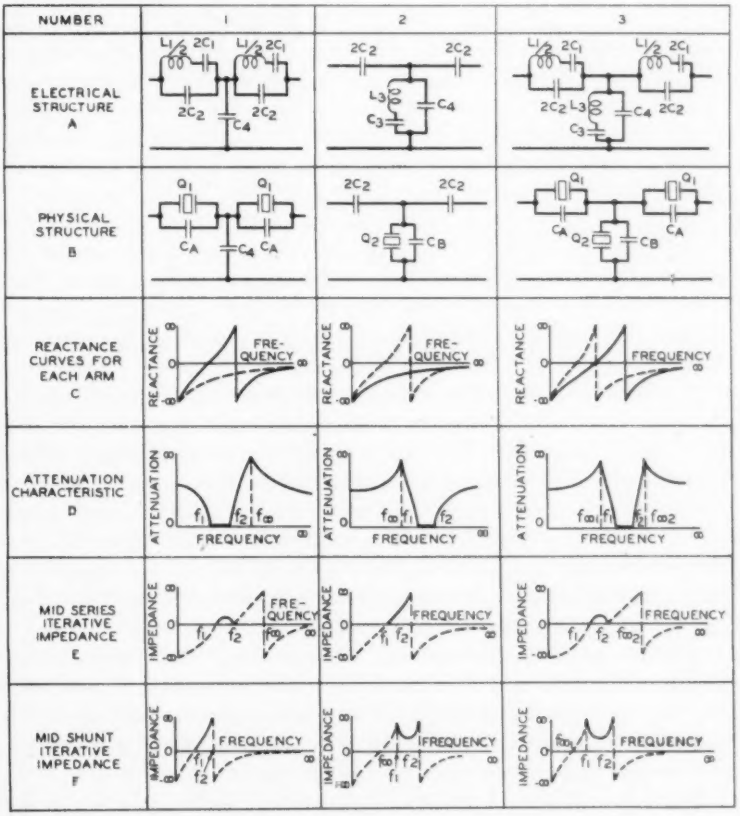


Fig. 6—Ladder networks employing crystals and condensers.

this type network, there are only three types of single band sections possible, all being band-pass filters. Figure 6 shows these sections, the impedance characteristic of each arm, the attenuation characteristics of these networks considered as filters, and their iterative impedances.

These attenuation characteristics and their limitations are at once found from a consideration of the impedance frequency curves for each arm shown by Fig. 6*C*. For a ladder type network it is well known [8] that a pass band will exist when

$$0 \leqq \frac{Z_1}{4Z_2} \leqq -1, \tag{5}$$

where Z_1 is the impedance of the series arm and Z_2 the impedance of the shunt arm. Hence, considering the first filter of Fig. 6, it is obvious that the lower cut-off fc_1 will come at the resonant frequency of the crystal. The upper cut-off fc_2 will come some place between the resonant and anti-resonant frequency, the exact position depending on the amount of capacitance in shunt. The anti-resonant frequency will be a point of infinite attenuation since the filter will have an infinite series impedance at this frequency.

With the restriction on the ratio of capacitances of the crystal noted in the previous section, it is easily shown that the ratio of the anti-resonant frequency to the resonant frequency is fixed and is about 1.004. Hence, we see that the ratio of $f\infty$ to fc_1 can be at most 0.4 per cent. The band width must be less than this since fc_2 must come between $f\infty$ and fc_1. A similar limitation occurs for the second filter of this figure, for which case the separation of $f\infty$ and fc_2 is at most 0.4 per cent. For filter number 3, a somewhat larger frequency separation between the points of infinite attenuation results, it being at most 0.8 per cent. The addition of any electrical capacitance in series or shunt with any of the crystals results in a narrowing of the band width.

It is seen then that there are two limitations in the types of filters obtainable with crystals and condensers in ladder sections. One, there is a limitation on the position of the peak frequencies and two, there is a limitation for the band width of the filters.

By employing the more general lattice type of filter section shown on Fig. 7, the first of these limitations can be removed. By means of this type of section it is possible to locate the attenuation peak frequencies at any position with respect to the pass band, but the pass band is limited in width to at most 0.8 per cent.

For a lattice filter a pass band exists when the impedances of the two arms are related by the expression [9]

$$0 \leqq \frac{Z_1}{Z_2} \leqq -\infty, \tag{6}$$

[8] See, for example, page 190 in book by K. S. Johnson, "Transmission Circuits for Telephonic Communications."

[9] "Physical Theory of the Electric Wave Filter," G. A. Campbell, *B. S. T. J.*, November, 1922.

where Z_1 is the impedance of the series arm (either 1, 2 or 3, 4 of Fig. 7*A*) and Z_2 the impedance of the lattice arm (either 1, 3 or 2, 4 of Fig. 7*A*). Hence, if one pair of branches has a reactance whose sign

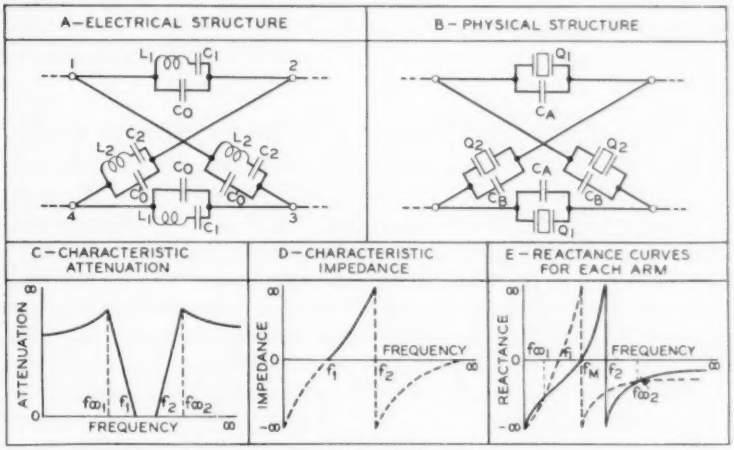


Fig. 7—Lattice network employing crystals and condensers.

is opposite to that of the other pair, a pass band exists, while if they are the same sign an attenuated band exists. Since the lattice is in the form of a Wheatstone bridge an infinite attenuation exists when the bridge is balanced, which occurs when both pairs of arms have the same impedance.

Let us consider a lattice filter with a crystal in each arm as shown by Fig. 7*B*. The crystals form two pairs of identical crystals, two alike in the series arms and two alike in the lattice arms. In order that a single band shall result it is necessary that the anti-resonant frequency of one arm coincide with the resonant frequency of the other as shown by Fig. 7*E*. It is obvious that the band width will be twice the width of the resonant region of the crystal or at most 0.8 per cent. Since the attenuation peaks occur when the two arms have the same impedance, they may be placed in any desired position by varying the impedance of one set of crystals with respect to the other. If crystals alone are used, these peaks will be symmetrical with respect to the pass band, but if in addition condensers are used with these crystals, the peaks may be made to occur dissymmetrically. In fact they may be made to occur so that both are on one side of the pass band. A narrower band results when capacitances are used in addition to crystals since the ratio of capacitances becomes larger. This may be utilized to control the width of the pass band to given any value less than 0.8 per cent.

The use of more crystals than four, in any network configuration employing only quartz crystals and condensers can be shown to result in no wider bands than 0.8 per cent, although higher losses can be obtained by the use of more crystals. Hence by the use of quartz crystals and condensers alone, a limitation in band width to 0.8 per cent is a necessary consequence of the fixed ratio of capacitances C_0/C_1 of equation (3).

Filter Sections Employing Crystals, Condensers and Coils

As was pointed out in the last section, filters employing only crystals and condensers are limited to band pass sections whose band widths do not exceed about 0.8 per cent. This band width is too narrow for a good many applications and hence it is desirable to obtain a filter section allowing wider bands while still maintaining the essential advantages resulting from the use of sharply resonant crystals. Such filters can be obtained only by the use of inductance coils as elements. Since the ratio of reactance to resistance of the best coils mounted in a reasonable space does not exceed 400, attention must be given to the effect of the dissipation.

The effect of dissipation in a filter is two-fold. It may add a constant loss to the insertion loss characteristic of the filter, and it may cause a loss varying with frequency in the transmitting band of the filter. The second effect is much more serious for most systems since an additive loss can be overcome by the use of vacuum tube amplifiers whereas the second effect limits the slope of the insertion loss frequency curve. Hence, if the dissipation in the coils needed to widen the band of the filter has only the effect of increasing the loss equally in the transmitting band and the attenuating band of the filter, a satisfactory result is obtained. The question is to find what configuration the coils must be placed in with respect to the crystals and condensers in order that their dissipation will not cause a loss varying appreciably with frequency.

Not every configuration will give this result, as is shown by the following example. The equivalent circuit of the crystal shown by Fig. 2*A* can be transformed into the form shown by Fig. 8*A* where the ratio $C_1/C_0 = 125$. This gives the same reactance curve as before, limited to a width of 0.4 per cent. Now suppose that we add an electrical anti-resonant circuit in series with the crystal—Fig. 8*B*—resonating at the same frequency and having the same constants as the anti-resonant network representing the crystal. If this circuit were dissipationless we could combine the two resonant circuits into one with twice the inductance and half the capacitance of that for the crystal alone

and hence the capacitance ratio would be 1/2 (125) or 62.5. The band width possible would then be twice that of the crystal alone. However, when the effect of dissipation is considered it is found that not much has been gained by employing the anti-resonant circuit. For the re-

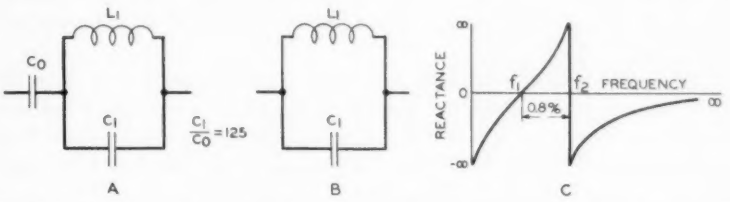


Fig. 8—Use of an anti-resonant circuit to broaden the resonance region of a crystal.

sistance, at resonance of the crystal and electrical circuit combination, will be the resistance of the electrical resonant circuit since that of the crystal is small compared to the electrical element. Hence we have doubled the impedance of the anti-resonant circuit and have the resistance of the electrical circuit. Hence the ratio (Q) of reactance to resistance of the anti-resonant circuit is double that of the electrical element alone. Even this Q, however, is insufficient to make a narrow band filter whose band width is 1.6 per cent (twice that possible with a crystal alone) and hence no useful purpose is served by combining a crystal with an electrical anti-resonant circuit.

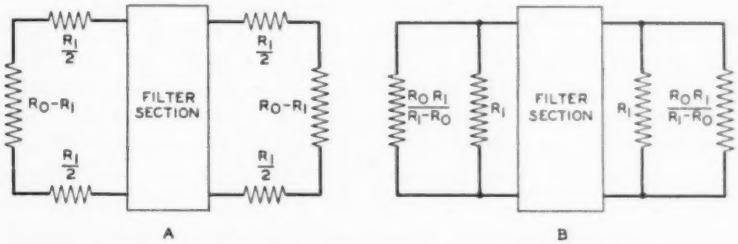


Fig. 9—Circuit showing resistances on the ends of filter sections.

Suppose, however, that all of the dissipation of the filter section be concentrated at the ends of the sections, either in series or in parallel with the filter as shown on Fig. 9. Then provided these resistances are within certain limits, they can be incorporated in the terminal resistances of the filter by making these resistances either smaller or larger for series or shunt filter resistances respectively. Between sections the resistances on the ends of the filter can be incorporated with other

resistances in such a way as to make a constant resistance attenuator of essentially the same impedance as the filter. For a series coil, this can be done by putting a shunt resistance between sections, while for a shunt coil it can be done by putting series resistances between sections. If this is done the whole effect of the dissipation is to add a constant loss to the dissipationless filter characteristic, this loss being independent of the frequency.

Since the lattice type network provides the most general type of filter network, attention will first be directed to this type of section employing inductances. It is easily proved that if any impedance is in series with both sides of a lattice network, as shown by Fig. 10*A*, then

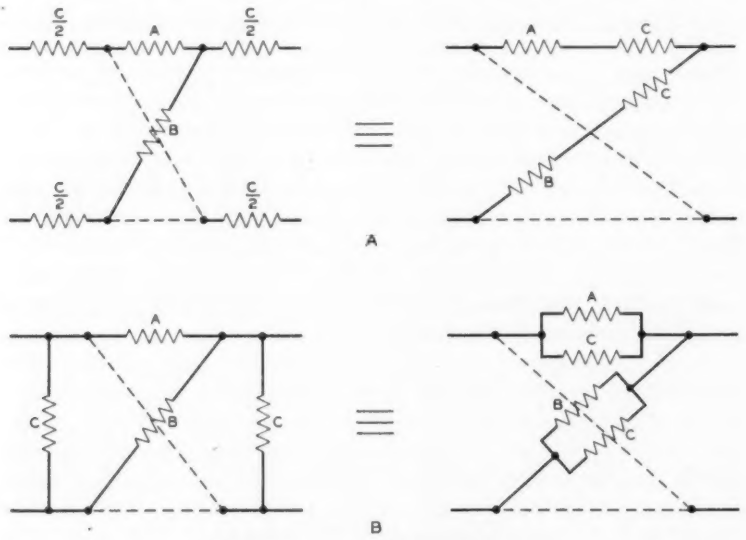


Fig. 10—Two network equivalences.

this is equivalent to placing this impedance in series with each arm of the lattice network as shown. Similarly, if a given impedance shunts the two ends of a lattice network, as shown by Fig. 10*B*, a lattice network equivalent to this is obtained by placing the impedance in shunt with all arms of the lattice. We are then led to consider a lattice network which contains coils either in series or in shunt with the arms of a lattice network, these arms containing only crystals and condensers, since the dissipation will then be effectively either in series or in shunt with the lattice network section.

If an inductance is added in series with a crystal the resulting re-

actance is shown by the full line of Fig. 11; the dotted lines show the reactance curves for the individual elements. It is evident that the resonant frequency of the crystal is lowered, the anti-resonant point remains the same, and an additional resonance is added at a frequency

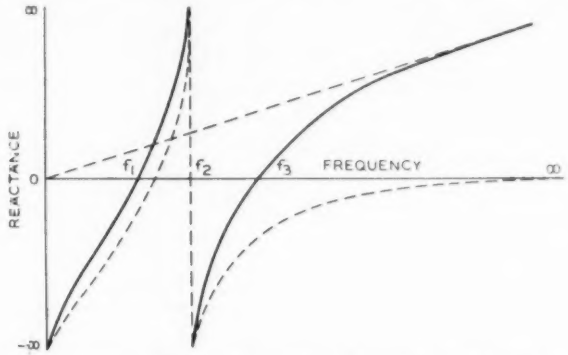


Fig. 11—Impedance characteristic of crystal and coil in series.

above the anti-resonant frequency. For a crystal whose ratio of capacitances r is about 125 it is easily shown by calculation that if the resonances are evenly spaced on either side of the anti-resonant frequency the percentage frequency separation between the upper resonance and the lower resonance is in the order of 9 per cent.

Suppose now that this element is placed in the series arm of a lattice network and another element of similar character is placed in the lattice arm, the second element having its lowest resonance coincide with the anti-resonance of the first element, and having the anti-resonance of the second element coincide with the highest resonance of the first element. This condition is shown by Fig. 12*C*. This network will produce a band-pass filter whose band extends from the lowest resonance of the series arm to the highest resonance of the lattice arm, a total percentage frequency band width of 13.5 per cent. By designing the impedances correctly the impedances of the two arms can be made to coincide three times so that there is a possibility of three attenuation peaks due to this section as shown by Fig. 12*D*. The loss introduced by the filter is equivalent to that introduced by three simple band-pass sections. Ordinarily the coils in the two arms are made equal so that their resistances are equal and for this case one of the peaks occurs at an infinite frequency. Since the resistances are equal, then by the theorem illustrated by Fig. 10*A* these resistances can be brought out on the ends and incorporated with the terminal

resistances, with the result that the dissipation of the coils needed to broaden the band has only the effect of adding a constant loss to the filter characteristic, this loss being independent of the frequency.

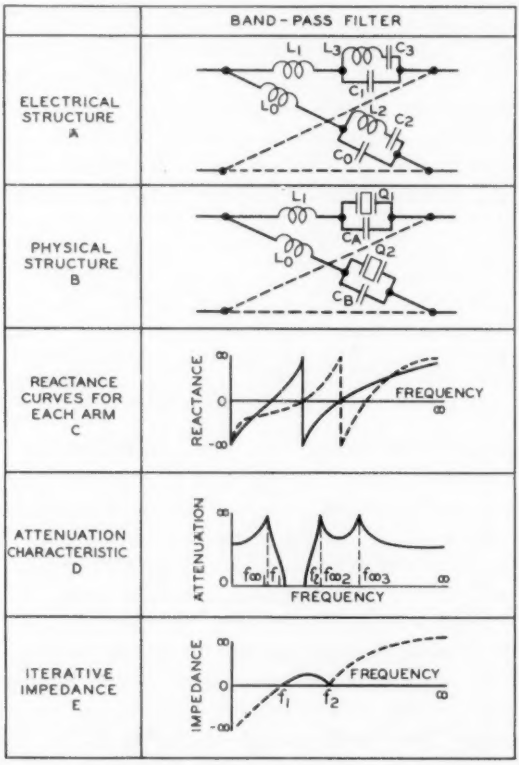


Fig. 12—Lattice network band-pass filter employing series coils.

To vary the width of the band below the 13.5 per cent band obtained with crystals only, added capacitances can be placed in parallel with the crystals increasing the ratio r. This results in a smaller separation in the resonant frequencies and hence a narrower band width. By this means the band width can be decreased indefinitely, although the dissipation caused by the coils introduces large losses for band widths much less than 1/2 per cent. By this means, however, it is possible to obtain band widths down to the widths which can be realized with crystals alone. On the upper side electrical filters can be built whose widths are as small as 13.5 per cent, hence this method fills

in a range not practical with electrical filters, or with crystals alone.

Another important characteristic of the filter is its iterative impedance. For a lattice filter this is given by [9]

$$Z_I = \sqrt{Z_1 Z_2},$$

where Z_1 is the impedance of the series arm and Z_2 that of the lattice arm. For a dissipationless filter, this is shown by Fig. 12*E*, as can be easily verified by a consideration of the reactance curves of Fig. 12*C*.

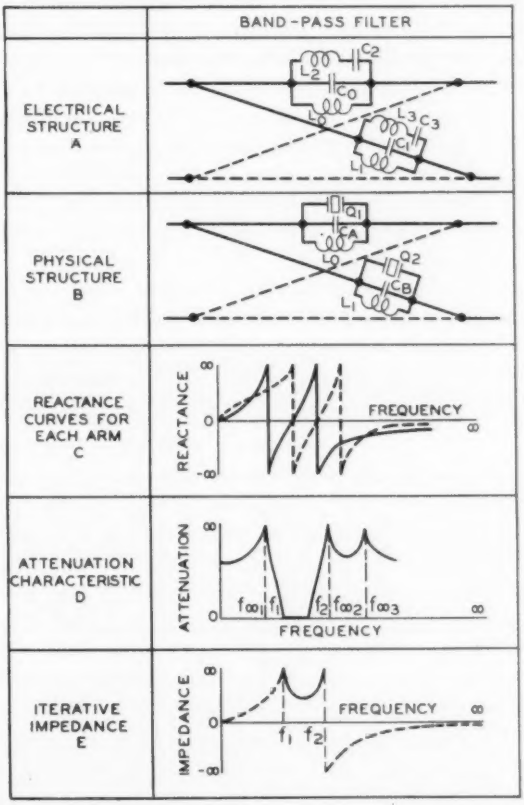


Fig. 13—Lattice network band-pass filter employing parallel coils.

This type of filter results in a relatively low impedance, for example about 600 ohms for a filter whose mid-band frequency is 64 kilocycles and whose band width is that shown on Fig. 19. Since the band width is decreased by adding more capacitance, it is evident that smaller

percentage band width filters will have lower impedances than the wider ones. For example, the filter whose characteristic is shown by Fig. 20, has an iterative impedance of 25 ohms.

It is evident that a still wider band can be obtained with the section discussed above by making the two resonances of Fig. 11 dissymmetrical. If the lower one is brought in closer to the anti-resonant frequency the top one extends farther out in such a manner that the total percentage frequency separation is greater than 9 per cent. If one element of this type is combined with one whose lower resonance is brought farther away from the anti-resonance than is the upper resonance, a filter whose pass band is greater than 13.5 per cent is readily obtained. On the other hand as the band is widened by this means, the cross-over points of the impedances of the two arms are of necessity brought very close to the cut-off frequencies, so that such a filter would introduce most of its loss very close to the cut-off frequencies. This type of characteristic might be useful in supplementing the loss characteristic possible with electrical elements, but by itself would not produce a very useful result.

We have so far discussed the characteristics which can be obtained by placing coils in series with crystals. An equally useful result is obtained by placing coils in shunt with crystals as shown by Fig. 13*B*. This arrangement results in a band-pass filter capable of giving the same band width as the first type discussed above. The only difference

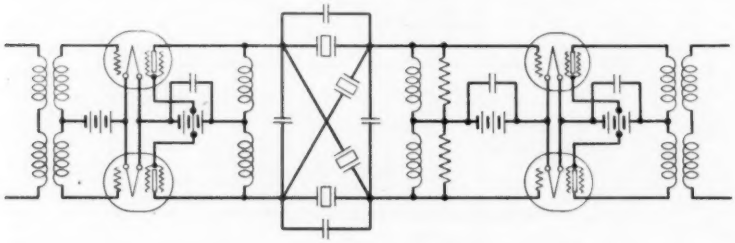


Fig. 14—Band-pass filter used between vacuum tubes.

occurs in the iterative impedance which will be as shown by Fig. 13*E*. For narrow band widths this type of filter has a very high iterative impedance. For example, for a one per cent band width, using ordinary sized coils and crystals, the iterative impedance may be as high as 400,000 ohms. Such filters can be used advantageously in coupling together high impedance screen gird tubes without the use of transformers. One such circuit is shown schematically by Fig. 14.

Filters made by using either series or shunt coils in conjunction

with condensers and crystals make very acceptable band-pass filters capable of moderate band widths. It is often desirable to obtain low and high-pass filters having a very sharp selectivity. The filter of Fig. 12 can be modified to give a high-pass characteristic by leaving out the coils in the series or lattice arms of the network. However, it will be found that the cross-over points in the impedance curve of necessity come very close to the pass band and hence no appreciable loss can be maintained at frequencies remote from the pass band. A broader and more useful characteristic is obtained by using a transformer having a preassigned coefficient of coupling, in conjunction with crystals and condensers, as the element for broadening the separation of resonances. Such an element is shown by Fig. 15*A*. As is well known, a trans-

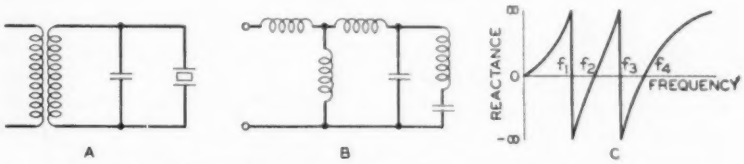


Fig. 15—Impedance characteristic of a transformer, condenser, and crystal.

former with a specified coupling can be replaced by a T network of three inductances as shown by Fig. 15*B*. The impedance characteristic, as shown by Fig. 15*C*, has two anti-resonant frequencies f_1 and f_3, and two resonant frequencies f_2 and f_4.

Suppose now that an element of this type is placed in one arm of the lattice and a similar element having a condenser in series with it is placed in the other arm as shown by Fig. 16*A*. If the elements are so proportioned that the anti-resonances of one arm coincide with the resonances of the other arm and vice versa, as shown by Fig. 16*B*, the impedances of the two arms are of opposite sign till the last resonance. Hence, a low pass filter results. It is possible to make the two impedance curves cross five times, so that an attenuation corresponding to five simple sections of low-pass filter results. Other arrangements of the resonances are also possible and are advantageous for special purposes. For example, as shown by Fig. 16*C*, we can make the last resonance and anti-resonance of both arms coincide, and the other resonances of one arm coincide with the anti-resonances of the second arm. This arrangement results in a low-pass filter having an attenuation corresponding to three simple low-pass filter sections and an impedance which can be made nearly constant to a frequency very near the cut-off frequency. This is advantageous for obtaining a filter with

a sharp cut-off, for otherwise the mismatch of impedance near the cut-off frequency causes large reflection losses which prevent the possibility of obtaining a sharp discrimination.

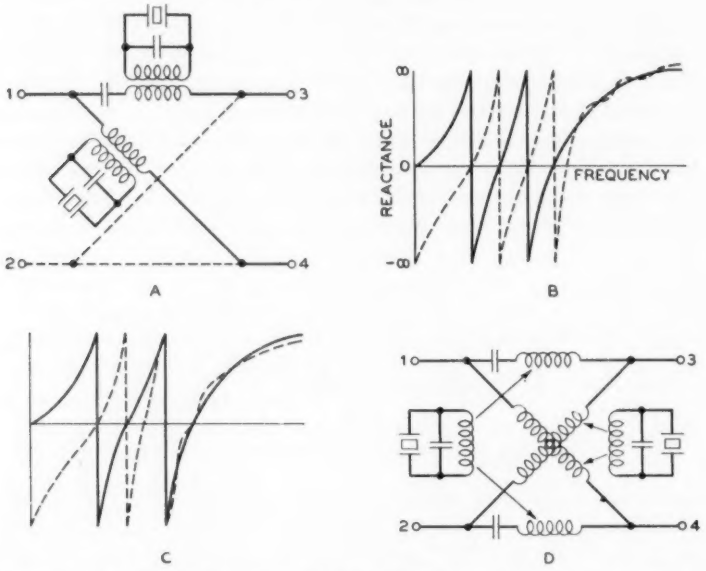


Fig. 16—Lattice network low-pass filter employing transformers, condensers, and crystals.

The effect of dissipation in the transformer on the loss characteristic is not so easy to analyze in this case as in the case of a series coil. The effect can be obtained approximately as follows. Of the three coils of Fig. 15*B* representing the transformer, the shunt coil has the least dissipation since no copper losses are included in this coil. For an air core coil, the Q of this shunt coil becomes very high and its dissipation can be neglected. The resistance of the primary winding can be incorporated in the terminal resistance as in the series coil type of filter and hence will cause only an added loss. The resistance of the secondary will be in series with the crystal and condenser, and for a reasonably good coil is of the same order of magnitude as the crystal resistance at resonance. Hence its effect will be much the same as cutting the Q of the crystal in half, so that instead of a crystal whose Q is 10,000, we use one whose Q is 5000 and a dissipationless coil. We see then that the Q of the crystal is still the most important factor in determining the sharpness of cut-off in the filter as in the previous

ones described, and hence a very sharp selectivity can be obtained with this circuit. It is possible to save elements in this filter by using two primaries for each coil, putting one primary in one series or lattice arm and the other in the corresponding series or lattice arms as shown by Fig. 16*D*. Only half the number of elements per section are required.

By replacing the series condenser of the series arm of Fig. 16*A* by a parallel condenser, it is possible to change the filter from a low-pass to a high-pass filter. Condensers in series, or in parallel with both arms result in wide band-pass filters. It is possible to obtain a wider pass band with this type of filter than with the single coil type since the resonances will be spread over a wider range of frequencies.

In a good many cases it is desirable to have unbalanced filter sections rather than the balanced type which results from the use of a lattice network. This is particularly true for high impedance circuits for use with vacuum tubes. Since the lattice type section is the most general type, it gives the most general characteristics obtainable. The filter sections described here can in some cases be reduced to unbalanced bridge T sections by well known network transformations, with, however, more restrictions on the type of attenuation characteristics physically obtainable.

A very simple bridge T network, which is equivalent to a lattice network of the kind shown on Fig. 13, with two crystals replaced by condensers, is shown on Fig. 17. This section employs mutual induct-

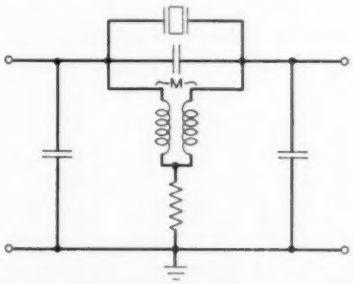


Fig. 17—Single crystal bridge T band-pass filter.

ance, and the resistance [10] shown is necessary in order to balance the arms of the equivalent lattice. This type of network is able to reproduce some of the characteristics of the lattice filter, but is not so general and is, moreover, affected by the dissipation of the coil to a larger extent than the equivalent lattice.

[10] The use of this resistance was suggested by Mr. S. Darlington and practically all the work of developing this filter has been done by Mr. R. A. Sykes.

Experimental Results

A number of filters have been constructed, during the past four years, which employ quartz crystals as elements. Figure 18 shows the measured insertion loss characteristic of a narrow band filter [11] employ-

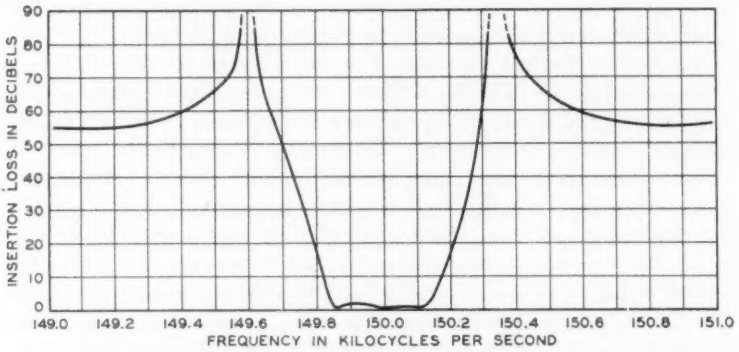


Fig. 18—Measured insertion loss characteristic of a narrow band-pass filter.

ing only crystals and condensers. This filter employs two sections of filter No. 3 of Fig. 6. It will be noted that in spite of the very narrow band width, the insertion loss in the transmitted band is quite small.

A number of the broader band filters employing coils as well as condensers and crystals have also been constructed. The frequency range so far developed extends from 36 kilocycles to 1200 kilocycles. Figure 19 shows the insertion loss characteristic of a band-pass filter whose mid-frequency is 64 kilocycles and whose band width is 2500 cycles. The insertion loss rises to 75 db, 1500 cycles on either side of the pass region. This filter was constructed from two sections of the band-pass type described in Fig. 12. A similar insertion loss characteristic, but shifted to a higher frequency, is shown by Fig. 20. The insertion loss in the center of the band for this higher frequency filter is considerably larger due to the smaller percentage band width. It is interesting to note that practically all of this loss is due to the dissipation introduced by the coils. The useful percentage band width is about one-half per cent and the filter reaches its maximum attenuation

[11] The filters whose characteristics are shown on Figs. 18 and 21 were designed and constructed by Messrs. C. E. Lane and W. G. Laskey. The author wishes to call attention to the fact that they and others associated with them in the Laboratories have made considerable progress in connection with the practical difficulties encountered in the design and construction of these filters such as working out the high precision element adjustment methods required, in methods of mounting, and in shielding methods.

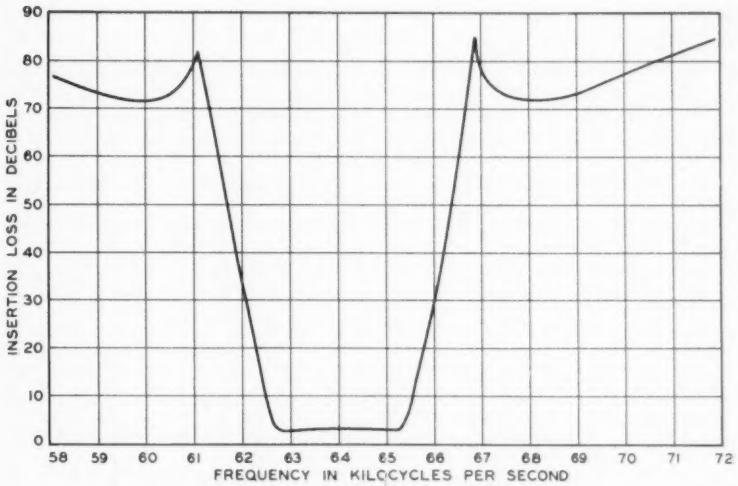


Fig. 19—Measured insertion loss characteristic of a band-pass filter.

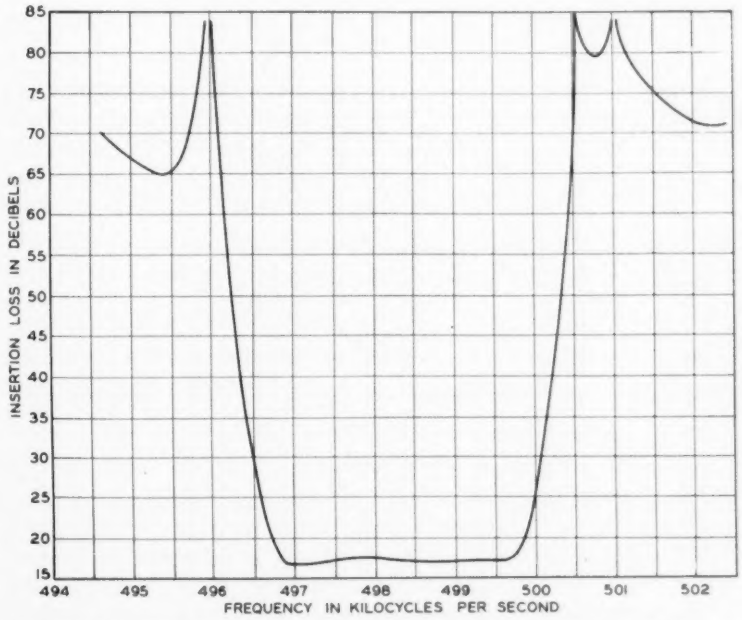


Fig. 20—Measured insertion loss characteristic of a band-pass filter at a high frequency.

in less than one-fourth per cent frequency range on either side of the pass band.

Figure 21 shows the insertion loss characteristics of a filter employed

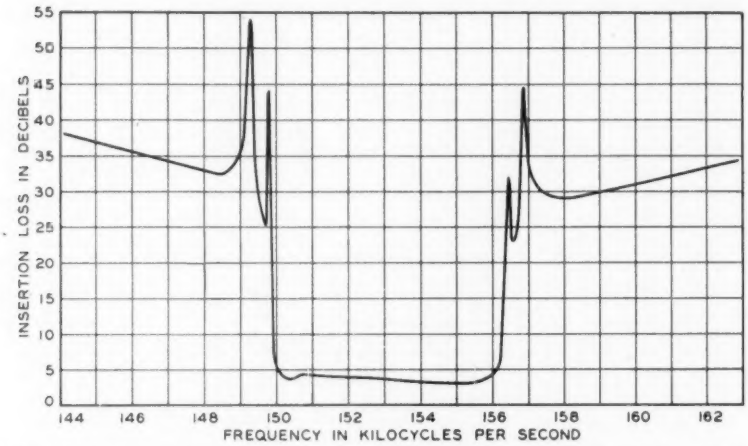


Fig. 21—Measured insertion loss characteristic of a band-pass filter used in a single side band radio receiver.

in an experimental radio system for separating the two sidebands of a channel at a high frequency. Here the separation is effected in about 0.15 per cent frequency range. With the best electrical filters the frequency space required for such a separation is about 1.5 per cent.

Figure 22 shows the insertion loss characteristic of a high-pass filter

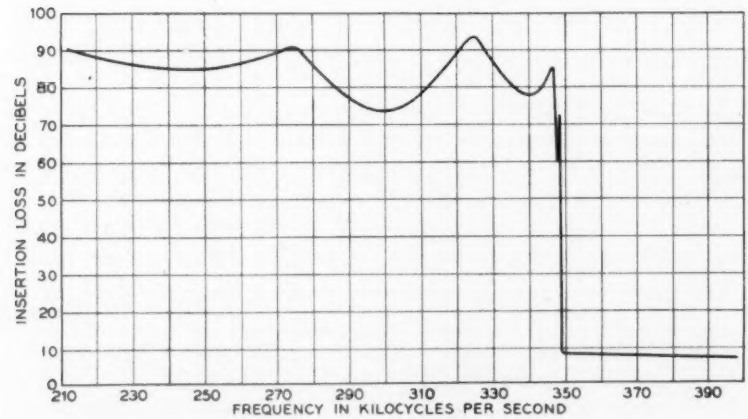


Fig. 22—Measured insertion loss characteristic of a high-pass filter.

constructed by using the circuit of Fig. 16*D*, modified by using a parallel condenser rather than a series condenser. The filter obtains a 65 db discrimination in less than a 0.12 per cent frequency separation.

Figure 23 shows a characteristic obtained by employing a filter of

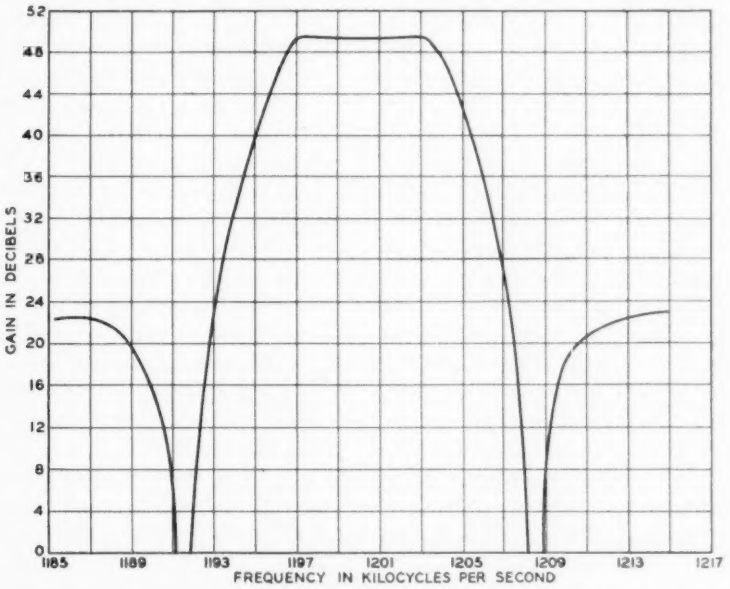


Fig. 23—Measured insertion loss characteristic of a single crystal bridge T filter.

the type shown by Fig. 17, together with a screen grid vacuum tube. The result is plotted as the gain of the circuit since this gives the most significant result for this type of circuit.

APPENDIX

The Modes of Vibration of a Perpendicularly Cut Quartz Crystal

Introduction

Quartz crystals have been cut into two principal types of orientations with respect to the natural crystal faces. The first type is the so-called Curie or perpendicular cut in which the crystal is so cut that its major surfaces are perpendicular to an electrical axis and parallel to the optical axis. Figure 1 shows such a cut. The second type is the so-called parallel or 30-degree cut in which the major surfaces are parallel

to both the optical and electrical axes. In this appendix a study is made of the modes of motion of a perpendicularly cut crystal. The effect has been studied of rotating the direction of the principal axis while still maintaining the principal surfaces perpendicular to the electrical axis. Such a crystal is designated as a perpendicularly cut crystal with an angle of rotation θ.

The perpendicularly cut crystal has received considerable theoretical and experimental consideration especially from Cady,[2] Van Dyke,[3] Dye [4] and Vigoreux.[5] They have assumed that the crystal has a plane wave vibration, and have calculated the frequencies of resonance in terms of the elastic constants and the density of the crystal, and have derived equivalent electrical networks for giving their electrical impedance. Such representations indicate that there should be one resonance for the crystal, the frequency of which is inversely proportional to the length and independent of the width of the crystal. As long as the length of the mechanical axis is large compared to that of any other axis, this prediction agrees with the experiment, but when the length of the other axes become comparable with that of the mechanical axis, the prediction is no longer fulfilled by experiment. It has long been recognized that this deviation is due to the failure of the plane wave assumption. Rayleigh [12] has given a correction for taking account of lateral motion, which is applicable to an isotropic medium. In a crystal, shear vibrations may be set up as well and for this case Rayleigh's correction can only be regarded as qualitative. Also if the other sets of resonance frequencies are to be investigated, account must be taken of the resonances of the other modes of vibration, and their reaction on the mode to be studied.

In this appendix experimental results have been obtained showing the frequencies of resonance found in perpendicularly cut crystals of various shapes and orientations. These frequencies are correlated with the elastic constants of the crystal and are shown to be completely accounted for by them. A coupled circuit representation is developed which is capable of predicting the main features of the principal vibration, including the change of frequency with the shape and orientation of the crystal, and the temperature coefficient curves.

Experimental Determination of the Resonant Frequencies

In order to investigate the modes of motion in a perpendicularly cut crystal in which the main axis coincides with the mechanical axis of the crystal, a set of measurements has been made on crystals whose

[2, 3, 4, 5] Loc. cit.
[12] Rayleigh, "Theory of Sound," Vol. I, Chapter VII, page 252.

mechanical axes are all 1.00 centimeter long, whose electrical axes are very thin, being 0.05 centimeter, and whose optical axes vary in dimension from 0.1 centimeter to 1.00 centimeter. In order to eliminate the effect of a series capacitance due to an air gap, the crystals were plated with a very thin coat of platinum. The effect of an added shunt capacitance in parallel with the crystal, due to the electrode capacitances, was practically eliminated by running the crystal electrodes in an outer grounded conductor as shown in Fig. 24, which shows the

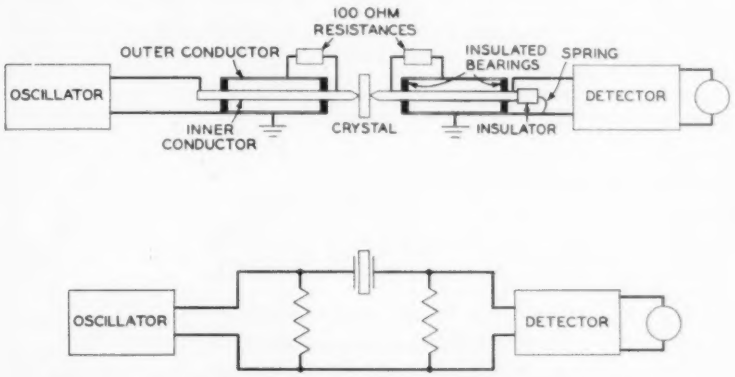


Fig. 24—Measuring circuit used to measure the resonances of a crystal.

measuring circuit. Contact to the crystal plating is made by means of small electrode points placed at the center of the crystal and kept in place by a small pressure. An increase in pressure over a moderate range was found to have no effect on the frequency of the crystal. The lowering of frequency due to plating was evaluated by depositing several films of known weight on the crystal and plotting its resonant frequency as a function of film weight. The intercept of this curve for a zero plating was taken as the frequency of the unplated crystal.

When the frequency of the oscillator was varied, the current in the detector showed frequencies of maximum and minimum current output which are respectively the frequencies of resonance and anti-resonance of the crystal. In order to locate accurately the frequencies of anti-resonance, it was found necessary to insert a stage of tuning in the detector, in order to discriminate against the harmonics of the oscillator. For a given crystal the frequency of the oscillator was varied over a wide range and the resonant and anti-resonant frequencies of the crystal were measured. The results of these measurements are shown by Fig. 25. In this curve the bottom part of the line repre-

sents the actual measured frequency, while the width of the line is proportional to the frequency difference between resonance and anti-resonance. In order to make this quantity observable, the frequency difference between resonance and anti-resonance is multiplied by a factor 6.

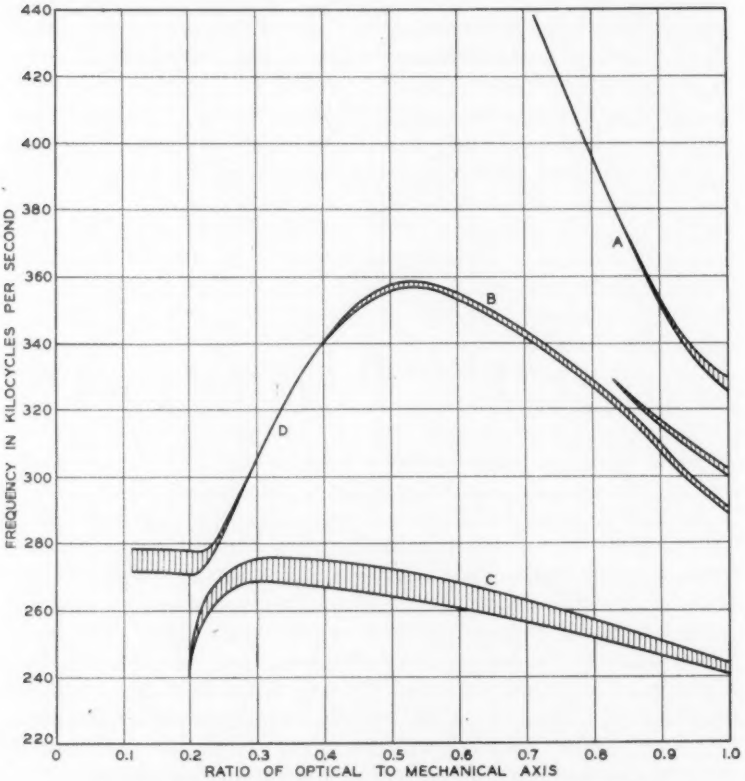


Fig. 25—Measured resonances of a perpendicularly cut crystal.

As long as the ratio of the optical to the mechanical axis is less than 0.2, the assumption of plane wave motion agrees well with experiment since there is only one resonance and its frequency does not depend to any great extent on the optical axis. However, above this point two frequencies make their appearance and react on each other to produce the coupled circuit curve shown. Finally when the ratio of optical to mechanical axes becomes larger a total of four resonant frequenices appear. Since a large number of crystals are used whose ratios of

optical to mechanical axes are greater than 0.2, it becomes a matter of some importance to investigate the causes of the additional resonances.

Interpretation of the Measured Resonance Frequency Curves of a Perpendicularly Cut Crystal

The plane wave assumption is valid for crystals whose width is less than 1/5 of their length, but it fails for wider crystals. It fails to represent a rectangular crystal because it does not allow for a wave motion in any other direction. That such a motion will occur is readily found by inspecting the stress-strain equations of a quartz crystal, given by equation (7).

$$\begin{aligned} -x_x &= s_{11}X_x + s_{12}Y_y + s_{13}Z_z + s_{14}Y_z, \\ -y_y &= s_{12}X_x + s_{11}Y_y + s_{13}Z_z - s_{14}Y_z, \\ -z_z &= s_{13}X_x + s_{13}Y_y + s_{33}Z_z, \\ -y_z &= s_{14}X_x - s_{14}Y_y + s_{44}Y_z, \\ -z_z &= s_{44}Z_x + s_{14}X_y, \\ -x_y &= s_{14}Z_x + \tfrac{1}{2}(s_{11} - s_{12})X_y, \end{aligned} \tag{7}$$

where x_x, y_y, z_z are the three components of extensional strain, and y_z, z_x, x_y the three components of shearing strains. X_x, Y_y, Z_z, Y_z, Z_x, and X_y are the applied stresses and s_{11}, etc. are the six elastic compliances of the crystal. Their values are not determined accurately but the best known values are given in equation (42). In this equation the X axis coincides with the electrical axis of the crystal, the Y axis with the mechanical axis, and the Z axis with the optical axis.

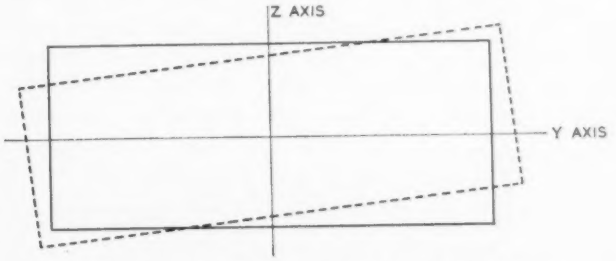


Fig. 26—Form of crystal distorted by an applied Y_y force.

Limiting ourselves now to an X or perpendicularly cut crystal the only stresses applied by the piezo-electric effect are an X_x, a Y_y, and a Y_z stress. Hence for such a crystal only four of the six possible types of motion are excited, three extensional motions x_x, y_y, z_z and one shear

motion y_z. Under static conditions, then, the motion at any point in the crystal is given as the sum of four elementary motions, three extensional motions and one shear motion. Moreover, these motions are coupled [13] as is shown by the fact that a force along one mode produces displacements in other modes of motion. Figure 26 shows how a perpendicularly cut crystal will be distorted for an applied Y_y force.

Suppose now that an alternating force is applied to the crystal. The simplest assumption that we can make regarding the motion is that the motion of any point is composed of four separate plane wave motions of the four types of vibration and that these react on each other in the way coupled vibrations are known to act in other mechanical [14] or electrical circuits. For the present purpose we can neglect motion along the X or electrical axis since this axis has been assumed small. The three remaining motions if existing alone will have resonances as shown by the solid lines of Fig. 27. That along the mechanical axis will have a constant frequency, since the mechanical axis is assumed constant, and is shown by the line C. The extensional motion along the optical axis will have a frequency inversely proportional to the length of the optical axis and will be represented by the line A of the figure. The shear vibration y_z, as shown by the section on the resonance frequency of a crystal vibrating in a shear mode, will have a frequency varying with dimension as shown by the line B.

In view of the coupling between the motions, the actual measured frequencies will be as shown by the dotted lines in agreement with well known coupled theory results.

If we compare these hypothetical curves with the actual measured values some degree of agreement is apparent. The main resonant frequency except in the region $0.2 < l_0/l_m < 0.3$ follows the dotted curve drawn. Also, the extensional motion along the optical axis has a frequency agreeing with that of Fig. 25. The shear vibration, however, has an entirely different curve from that conjectured. What is happen-

[13] The idea of elementary motions in the crystal being coupled together appears to have been first suggested in a paper by Lack "Observations on Modes of Vibration and Temperature Coefficients of Quartz Crystal Plates," *B. S. T. J.*, July, 1929, and was used by him to explain the effect of one mode of motion on the temperature coefficient of another mode and vice versa. The idea of associating this coupling with the elastic constants of the crystal occurred to the writer in 1930 but was not published at that time. It is, however, incorporated in a patent applied for some time ago on the advantages of crystals cut at certain orientations. More recently the same idea is given in a paper by E. Giebe and E. Bleckschmidt, *Annalen der Physik*, Oct. 16, 1933, Vol. 18, No. 4. They have extended their numerical calculations to include three modes of motion.

[14] This coupling is shown clearly for a mechanical system by one of the few rigorously solved cases of mechanical motion for two degrees of freedom—the vibration of a thin cylindrical shell—given by Love in "The Mathematical Theory of Elasticity," Fourth Edition, page 546.

ing there is I think evident from a consideration of Fig. 28. Here in solid lines are drawn two frequency curves one of which, *B*, is the shear frequency curve of Fig. 27. The other curve, *D*, has a rising frequency with an increase in the optical axis dimension. Assuming these vibra-

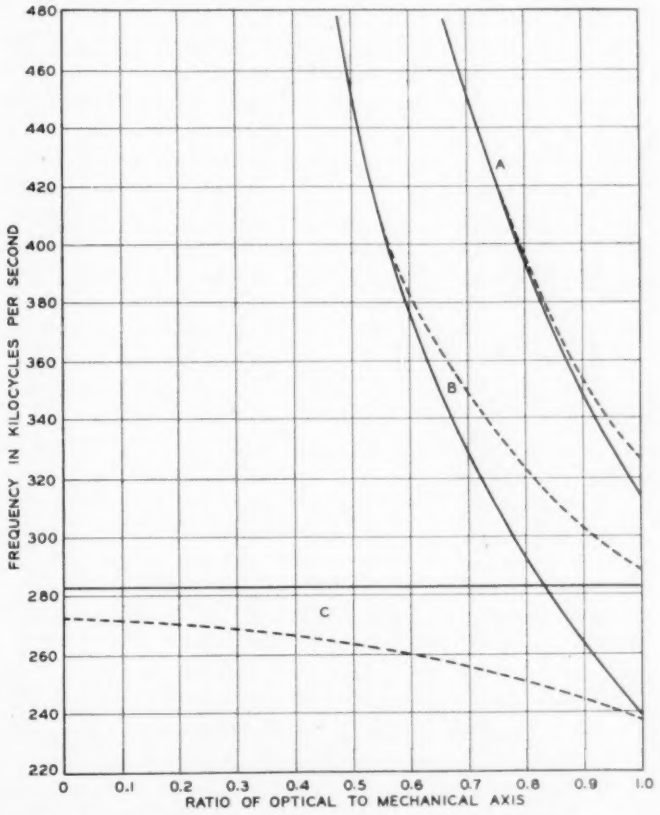


Fig. 27—Theoretical resonances of a perpendicularly cut crystal showing effect of coupling.

tions coupled a resonance frequency curve shown by the dotted line will be obtained. If this curve is substituted for the shear curve of Fig. 27 and the actual resonant frequency raised to take account of the effect of coupling with the longitudinal motion along the mechanical axis, a curve very similar to the measured curve of Fig. 25 is obtained.

The type of motion coupled to the shear motion is easily found. Its

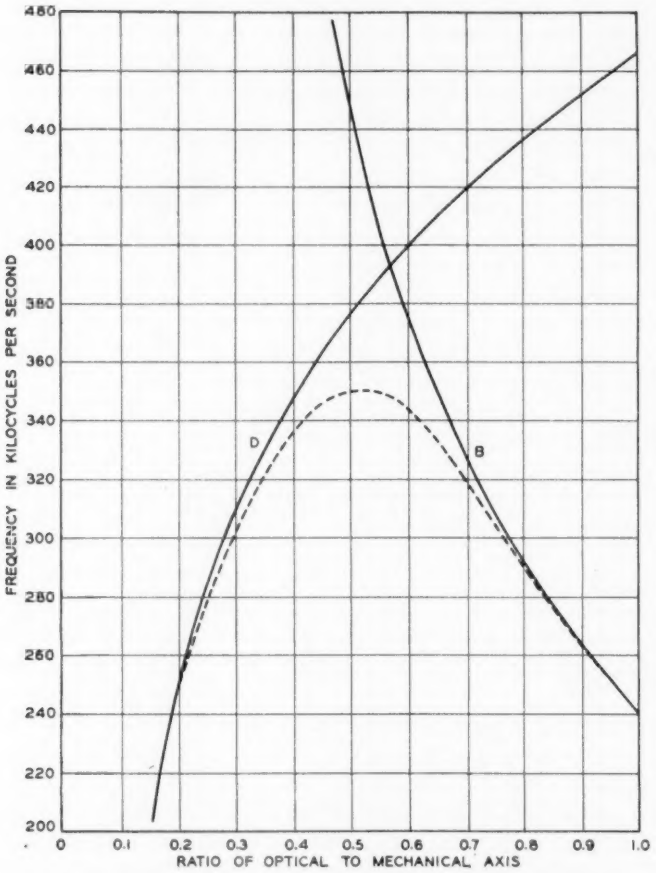


Fig. 28—Coupled frequency curve for shear and flexure vibrations.

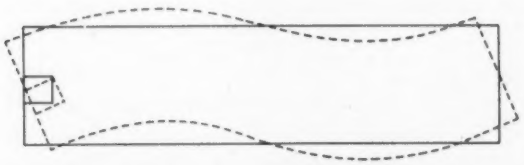


Fig. 29—Bar bent in its second flexural mode of vibration.

frequency increases as the optical axis dimension is increased and about the only type of motion which does this is a flexural motion as shown by Fig. 29. This figure shows the second type of motion possible to a bar in flexure rather than the first for experiments by Harrison [15] show that the frequency for the first type of motion is too low to account for this vibration. Harrison has also measured the frequencies of a bar in its second flexural mode and the solid line, D, of Fig. 28 is an actual plot of these measured frequencies up to a ratio of $l_0/l_m = 0.25$, which is as far as Harrison carries his measurements. The rest of the curve is obtained by extrapolation. There is no doubt then that a flexural motion is involved in this coupling. The mechanism by which the bar is driven in flexure will be evident if we observe what happens to a square on the crystal in the unstrained state. As shown by Fig. 29, its deformation is similar to that of a shear deformation. The amount of shear depends on the distance from the nodes of the crystal. Some of the shear is in one direction and some in the other but the two amounts are not balanced and hence a pure shear in one direction can excite a flexural motion of the crystal.

The strength of the coupling from the mechanical axis motion y_y to the shear motion y_z and the extensional motion along the optical axis z_z are indicated by the coupling compliances $s_{24}/\sqrt{s_{22}s_{44}}$ and $s_{23}/\sqrt{s_{22}s_{33}}$, respectively. From the values of these constants we find that the shear motion is more closely coupled than the z extensional motion, and this is indicated experimentally by the greater width of the shear line.

Effect of a Rotation of the Longest Axis with Respect to the Electrical Axis on the Resonances of a Crystal

From the qualitative explanation of the secondary resonances given above, it is possible to predict how these resonances will be affected by any change in the crystal which changes the constants determining the three modes of motion and their coupling coefficients. One method for varying these constants is to change the direction for cutting the crystal slab from the natural crystal. In the present paper consideration is limited to those crystals which have their major faces perpendicular to an electrical axis, i.e., a perpendicularly cut crystal with its longest direction rotated by an angle θ from the direction of the mechanical axis. The convention is adopted that a positive angle is a clockwise rotation of the principal axis for a right handed crystal, when the electrically positive face (determined by a squeeze) is up.

[15] "Piezo-Electric Resonance and Oscillatory Phenomena with Flexural Vibration in Quartz Plates," J. R. Harrison, *I. R. E.*, December, 1927.

For a left-handed crystal a positive angle is in a counter-clockwise direction.

In the section dealing with elastic and piezo-electric constants for rotated crystals (page 449) is given a method for determining the elastic constants of a rotated crystal and curves are given for the ten elastic constants. These have been worked out by Mr. R. A. Sykes of the Laboratories. The method of designation is the following: The X axis remains fixed and is designated by $1'$. The axis of greatest length is designated by $2'$, since in the unrotated crystal the mechanical axis, corresponding to the Y direction, is the axis of greatest length. Extensional motion perpendicular to the $2'$ axis is designanted by $3'$, and shear motion in the plane determined by the $2'$, $3'$ axes is designated by $4'$. The ten resulting constants s_{11}', s_{22}', s_{33}', s_{44}', s_{12}', s_{13}', s_{14}' s_{23}', s_{24}', s_{34}' are shown evaluated in terms of the angle θ on Fig. 30. Since

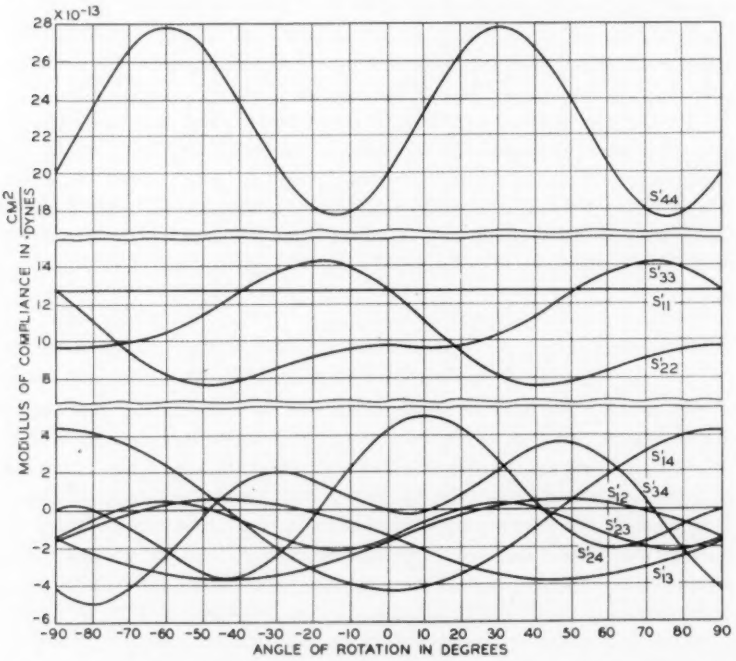


Fig. 30—Elastic compliances of a perpendicularly cut crystal as a function of the angle of rotation.

motion and coupling to motion along the X axis can be neglected, the constants of interest are s_{22}', s_{33}', s_{44}', s_{23}', s_{24}', s_{34}'. Since the $2'$ or y'

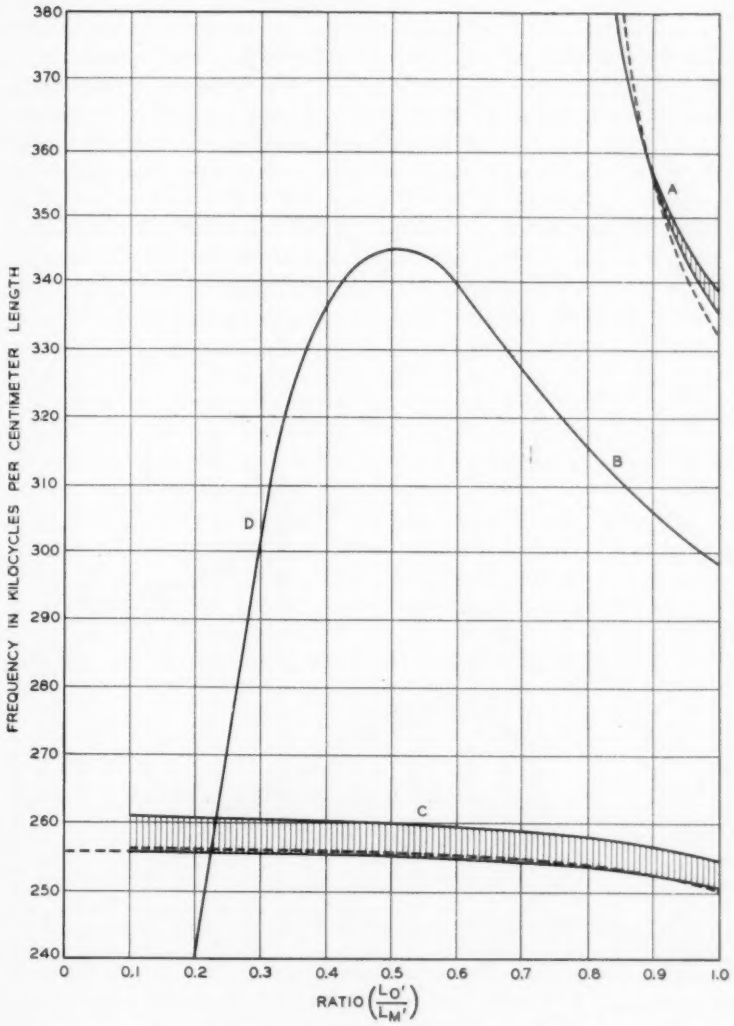


Fig. 31—Measured resonances of a $\theta = -18.5°$ perpendicularly cut crystal.

axis is the principal axis of motion, the mutual compliances of principal interest are s_{23}', determining the coupling between the Y' extensional motion and the Z' extensional motion, and s_{24}' determining the coupling between the Y' extensional motion and the Y_z' shear motion. It is the shear motion which is most objectionable, because it is more highly coupled than the Z' extensional motion, because it is lower in frequency, and because it is coupled to a flexural mode. Hence, if this motion can be eliminated or made very small, a much better crystal for most purposes is obtained. We note that if θ is $-18.5°$ or if $\theta = 41.5°$ the shear coupling coefficient vanishes and hence a force in the Y_y direction produces no y_z shear or vice versa. Of these the $-18.5°$ crystal is driven more strongly by the piezo-electric effect and hence has a more prominent resonance.

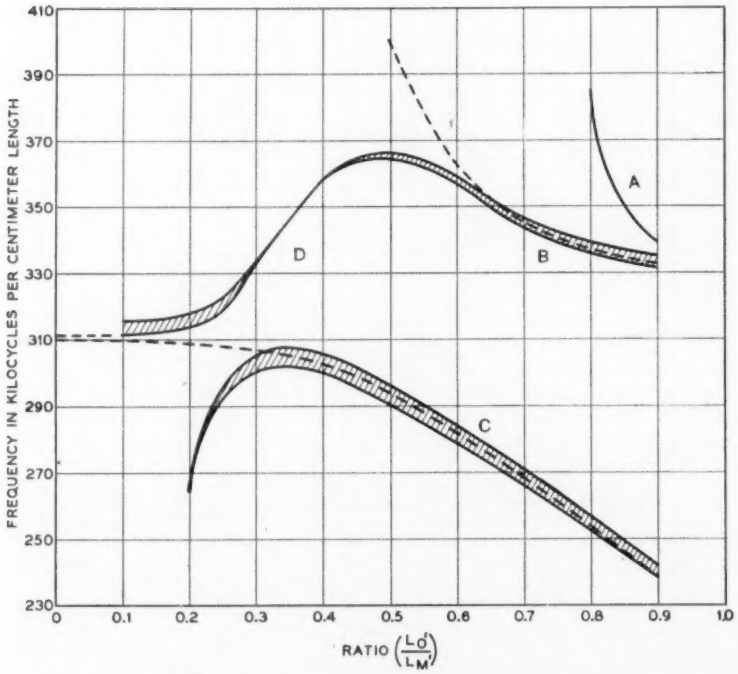


Fig. 32—Measured resonances of a $\theta = +18.5°$ perpendicularly cut crystal.

Accordingly the resonances of a $\theta = -18.5°$ cut crystal have been measured in a similar way to the $\theta = 0°$ cut crystal shown in Fig. 25. The result is shown on Fig. 31. As will be seen from the figure, the

shear resonance indicated by B is barely noticeable, while the z extensional mode indicated by A is somewhat stronger although higher in frequency. The frequency of the principal mode is not greatly affected by an increase in the z' axis until the ratio of axes is greater than .6. Another angle of some interest is $\theta = +18.5°$ since there the z' extensional coupling disappears. The resulting resonances are shown on Fig. 32. It will be noted that the z' extensional resonance curve A is very weak, while the shear curve B is quite pronounced.

An Equivalent Electrical Circuit for a Crystal Possessing Two Degrees of Motion

The above explanation accounts qualitatively for all the resonances observed in the crystal and how they are varied by a rotation of the crystal. It is desirable, however, to see if a quantitative check can be obtained from the known elastic constants of the crystal. To obtain a complete check would require a system capable of five degrees of motion. However, if we take the simplest case, the -18.5 degree cut crystal, only two modes of motion have to be considered, and even for the zero cut crystal, a good agreement is obtained by lumping the shear and extensional mode as one mode of motion and considering its reaction on the fundamental mode. Hence consideration is limited in this paper to a circuit having two modes of motion.

The properties of a single mode of motion can be represented for frequencies which do not exceed the first resonant frequency of the crystal, by the simple electrical circuit of Fig. 33*A*. Here the capaci-

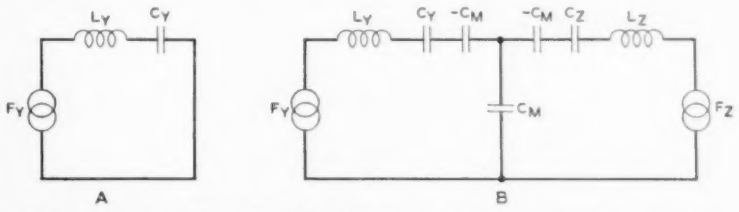


Fig. 33—Equivalent electrical circuit of a crystal having two modes of motion.

tance represents the mechanical compliance of the bar, the charge on the condenser represents a displacement per unit length of the bar, while the current flowing through the circuit represents the velocity of a point on the bar. The inductance represents the mass reaction of the crystal. The representation of the motion of a bar by a simple lumped circuit assumes that the bar moves as a whole, that is, if a force is applied to the body it contracts or expands equally at all parts of the

bar. This is contrary to actual conditions, since expansions or contractions proceed in the form of a wave from the ends of the bar toward the center. However, if consideration is limited to low frequencies, i.e. frequencies which do not exceed by much the first resonance of the bar, the approximation is good and a considerable simplification in the analysis is made. To take account of wave motion, the representation has to be an electric line as was pointed out in connection with acoustic filters.[16]

To represent two separate modes of motion and their coupling, the circuit shown by Fig. 33*B* is employed. A little consideration shows that the type of coupling existing in a crystal is capacitative since an extension along the mechanical axis produces a contraction along the optical axis, and vice versa. Since strains in mechanical terms are equivalent to charges in electrical terms, this type of coupling can be represented only by a capacitative network. This representation is entirely analogous to the T network representation for a transformer.[17] The constants of the network can be evaluated in terms of the elastic constants of the crystal as follows: For a -18.5 degree cut crystal, we can write the stress strain equation (7) as

$$\begin{aligned} y_y &= s_{22}' Y_y + s_{23}' Z_z, \\ z_z &= s_{23}' Y_y + s_{33}' Z_z, \end{aligned} \tag{8}$$

since we are neglecting motion along the X axis and since s_{24}', the coupling coefficient of the shear to the Y' axis is zero. No Y_z force is assumed acting. If we work out the equation for the charges on the condensers of the equivalent representation shown in Fig. 33*B* we have, with the charges and voltages directed as shown

$$\begin{aligned} Q_1 &= \frac{e_y C_y}{1 - K^2} + e_z \frac{\sqrt{C_y C_z} K}{1 - K^2}, \\ Q_2 &= \frac{e_y \sqrt{C_y C_z} K}{1 - K^2} + \frac{e_z C_z}{1 - K^2}, \end{aligned} \tag{9}$$

where K the coupling factor between the two modes of motion, is defined by the relation

$$K = \frac{\sqrt{C_y C_z}}{C_m}. \tag{10}$$

Associating Q_1 with y_y, the displacement per unit length, Q_2 with

[16] See "Regular Combination of Acoustic Elements," W. P. Mason, *B. S. T. J.*, April, 1927, p. 258.

[17] See, for example, p. 281 in the book "Transmission Circuits for Telephonic Communication" by K. S. Johnson.

z_x, e_y with Y_y and e_z with Z_z, we have on comparing (9) with (8)

$$s_{22}' = \frac{C_y}{1 - K^2}\,;\; s_{23}' = \frac{\sqrt{C_y C_z}K}{1 - K^2}\,;\; s_{33}' = \frac{C_z}{1 - K^2} \tag{11}$$

or inversely

$$C_y = s_{22}'\left[1 - \frac{s_{23}'^2}{s_{22}' s_{33}'}\right];\; C_z = s_{33}'\left[1 - \frac{s_{23}'^2}{s_{22}' s_{33}'}\right];\; K = \frac{s_{23}'}{\sqrt{s_{22}' s_{33}'}}. \tag{12}$$

If, now, alternating forces are applied to the crystal, another reaction to the applied force enters, namely the mass reaction of the crystal due to the inertia of the different parts. To take account of this reaction, the inductances are added to the two meshes representing mass reaction for the two modes of motion. To determine the value of the inductance, consider first the representation for one mode of motion shown by Fig. 33*A*. The resonant frequency of the system is given by

$$f_r = \frac{1}{2\pi\sqrt{LC}}. \tag{13}$$

On the other hand, the resonant frequency of a bar is given by

$$f_r = \frac{1}{2l\sqrt{\rho s}}, \tag{14}$$

where l is the length of the bar, s its compliance, and ρ its density. But in the above representation the capacitance C is the compliance constant s so that, on comparing (13) and (14) we find

$$L = \frac{l^2\rho}{\pi^2}. \tag{15}$$

In a similar manner for the coupled circuit, Fig. 33*B*, there results

$$L_y = \frac{l_y^2\rho}{\pi^2}\,;\; L_z = \frac{l_z^2\rho}{\pi^2}, \tag{16}$$

where l_y is the length of the crystal in centimeters along the y axis, and l_z the length of the crystal in centimeters along the z axis. Hence all of the constants of Fig. 33*B*, which represents the crystal for mechanical vibrations subject to the restrictions noted above, are determined and we should be able to predict all of the quantities which depend only on the mechanical constants of the crystal.

Of these the most important are the resonance frequencies of the crystal and their dependence on dimension, temperature coefficient and the like. To determine the natural mechanical resonance of a crystal,

we solve the network of Fig. 33*B* to find the frequencies of zero impedance for either an applied Y_y force or an applied Z_z force. The result is two frequencies f_1 and f_2 given by the coupled circuit equations

$$\begin{aligned} f_1^2 &= \tfrac{1}{2}[(f_A^2 + f_B^2) - \sqrt{(f_B^2 - f_A^2)^2 + 4K^2 f_A^2 f_B^2}], \\ f_2^2 &= \tfrac{1}{2}[(f_A^2 + f_B^2) + \sqrt{(f_B^2 - f_A^2)^2 + 4K^2 f_A^2 f_B^2}], \end{aligned} \tag{17}$$

where

$$f_A = \frac{1}{2\pi\sqrt{L_y C_y}} \text{ and } f_B = \frac{1}{2\pi\sqrt{L_z C_z}}. \tag{18}$$

Then f_A and f_B represent the natural frequencies along the Y and Z axis respectively when these two motions are not coupled together.

Two limiting cases of interest are obtainable from these relations. If f_B is much larger than f_A, the equations reduce to

$$\begin{aligned} f_1 &= f_A\sqrt{1 - K^2} = \frac{1}{2l_y\sqrt{\rho s_{22}'}}, \\ f_2 &= f_B = \frac{1}{2l_z\sqrt{\rho s_{33}'[1 - s_{23}'^2/s_{22}' s_{33}']}} \end{aligned} \tag{19}$$

upon substituting the value of the constants given before. The first equation shows that for a long thin rod the frequency depends on the elastic constant s_{22}', which is the inverse of Young's modulus. For the frequency f_2, which corresponds to that of a thin plate, a different elastic constant appears. Upon evaluating the expression $s_{33}'[1 - s_{23}'^2/s_{22}' s_{33}']$ in terms of the elastic constants which express the forces in terms of the strains—see equation (25)—we find that $s_{33}'(1 - s_{23}'^2/s_{22}' s_{33}') = 1/c_{33}$. c_{33} measures the ratio of force to strain when all the other coupling coefficients are set equal to zero, and corresponds to the frequency of one mode vibrating by itself without coupling to other modes. Hence the frequency of a thin plate should be

$$f = \frac{1}{2t}\sqrt{\frac{c_{nn}}{\rho}}, \tag{20}$$

where c_{nn} represents the elastic coefficient for the mode of motion considered, and t is the thickness of the plate. This deduction has been verified by experimental tests on thin plates.

Let us consider now the curves for the -18.5 degree cut crystal shown by Fig. 31. The values of the elastic constants for this case are

$$s_{22}' = 144 \times 10^{-14} \text{ cm.}^2/\text{dynes}; \; s_{23}' = -21.0 \times 10^{-14}; \; s_{33}' = 92.5 \times 10^{-14}. \tag{21}$$

Hence from equations (17), (18) and (21) one should be able to check the measured frequency curves of Fig. 31. The result is shown on the dotted lines of these curves. The agreement is quite good although a slightly better agreement would be obtained if s_{23} had a smaller value. Since these constants have never been measured with great accuracy, it is possible that they deviate somewhat from the curves of Fig. 30.

This theory can be applied also to a $\theta = +18.5$ degree cut crystal since the extensional coupling coefficient vanishes for this angle. The agreement is quite good if the frequency for the uncoupled mode given by the section on vibration in shear mode (page 446) is used in place of equation (14). The resonances for the $\theta = 0°$ cut crystal shown by Fig. 25 cannot be accounted for quantitatively by the simple theory given here since there are three modes of motion operating. The shear mode of motion is more closely coupled to the principal mode than is the Z_s extensional mode and hence a fair approximation is obtained by considering only the shear mode. However, for complete agreement the theory should be extended to a triply coupled circuit and that is not done in this paper.

Another phenomenon of interest which can be accounted for by the circuit of Fig. 33*B* is the temperature coefficient of the crystal and its variation with different ratios of axes and different angles of rotation. To obtain the relation, we assume that each of the vibrations may have a temperature coefficient of its own as may also the coefficient of coupling K. If a small change of temperature occurs, f_1 will change to $f_1(1 + T\Delta T)$, f_A to $f_A(1 + T_A\Delta T)$, f_B to $f_B(1 + T_B\Delta T)$ and K to $K(1 + T_K\Delta T)$. Assuming ΔT small so that its squares and higher powers can be neglected, we find from equation (17) that

$$T = \frac{1}{2f_1^2}\left[T_A f_A^2 \left[1 + \frac{f_B^2(1 - 2K^2) - f_A^2}{\sqrt{(f_B^2 - f_A^2)^2 + 4K^2 f_A^2 f_B^2}} \right] + T_B f_B^2 \right.$$
$$\times \left[1 - \frac{f_B^2(1 + 2K^2) - f_A^2}{\sqrt{(f_B^2 - f_A^2)^2 + 4K^2 f_A^2 f_B^2}} \right]$$
$$\left. - \frac{2T_K f_A^2 f_B^2}{\sqrt{(f_B^2 - f_A^2)^2 + 4K^2 f_A^2 f_B^2}} \right]. \quad (22)$$

The temperature coefficients of the six elastic constants have been measured at the Laboratories [18] by measuring the frequency temperature coefficients of variously oriented crystals. The temperature coefficients of the six elastic constants can be calculated from these

[18] These coefficients have been evaluated in cooperation with Messrs. F. R. Lack, G. W. Willard and I. E. Fair and their work is discussed in detail in their companion paper in this issue of the *B. S. T. J.*

measurements and have been found to be, in parts per million per degree centigrade:

$$T_{s_{11}} = +13;\ T_{s_{12}} = -1230;\ T_{s_{13}} = -347; \tag{23}$$
$$T_{s_{14}} = +130;\ T_{s_{33}} = +213;\ T_{s_{44}} = +172.$$

Using these values and neglecting the extensional motion Z_z, the temperature coefficients calculated from equation (22) for a 0 degree cut crystal are shown on the dotted line of Fig. 5 and agree quite well with the measured values.

The Resonance Frequencies of a Crystal Vibrating in a Shear Mode

The equations of motion for any aelotropic body are

$$\rho\frac{\partial^2 u}{\partial t^2} = \frac{\partial X_x}{\partial x} + \frac{\partial X_y}{\partial y} + \frac{\partial X_z}{\partial z},$$
$$\rho\frac{\partial^2 v}{\partial t^2} = \frac{\partial Y_x}{\partial x} + \frac{\partial Y_y}{\partial y} + \frac{\partial Y_z}{\partial z}, \tag{24}$$
$$\rho\frac{\partial^2 w}{\partial t^2} = \frac{\partial Z_x}{\partial x} + \frac{\partial Z_y}{\partial y} + \frac{\partial Z_z}{\partial z},$$

where u, v, w are the displacements of any point in the crystal along the x, y, z axes respectively and X_x, etc. are the six applied stresses. The strains have been expressed in terms of the stresses by equation (7). It is more advantageous for the present purpose to express the stresses in terms of the strains, which can be done by the following equations:

$$\begin{aligned} X_x &= c_{11}x_x + c_{12}y_y + c_{13}z_z + c_{14}y_z, \\ Y_y &= c_{12}x_x + c_{22}y_y + c_{23}z_z + c_{24}y_z, \\ Z_z &= c_{13}x_x + c_{23}y_y + c_{33}z_z, \\ Y_z &= c_{14}x_x + c_{24}y_y + c_{44}y_z, \\ Z_x &= c_{44}z_x + c_{14}x_y, \\ X_y &= c_{14}z_x + \tfrac{1}{2}(c_{11} - c_{12})x_y, \end{aligned} \tag{25}$$

where the c's are the elastic constants and the strains x_x, etc., are given in terms of the displacements u, v, w by the equations

$$x_x = \frac{\partial u}{\partial x};\ y_y = \frac{\partial v}{\partial y};\ z_z = \frac{\partial w}{\partial z};\ y_z = \left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}\right);$$
$$z_x = \left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}\right);\ x_y = \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right). \tag{26}$$

In equation (24) there exist the reciprocal relations

$$X_y = Y_x;\ X_z = Z_x;\ Y_z = Z_y. \tag{27}$$

For a free edge, i.e. no resulting forces being applied to the crystal, the conditions existing for every point of the boundaries are

$$\begin{aligned} X_\nu &= X_x \cos(\nu,x) + X_y \cos(\nu,y) + X_z \cos(\nu,z) = 0, \\ Y_\nu &= Y_x \cos(\nu,x) + Y_y \cos(\nu,y) + Y_z \cos(\nu,z) = 0, \\ Z_\nu &= Z_x \cos(\nu,x) + Z_y \cos(\nu,y) + Z_z \cos(\nu,z) = 0, \end{aligned} \tag{28}$$

where ν is the normal to the boundary under consideration.

If these equations are combined and completely solved, the motion of a quartz crystal is completely determined. The results which were obtained above in an approximate manner could be rigorously solved. However, on account of the difficulty [19] of the solution, this is not attempted here. In the present section it is simply the purpose to find out what resonances a crystal will have if it is vibrating in a shear mode only. To avoid setting up motion in the other modes of vibration, the coupling elasticities c_{14}, c_{24}, c_{34} are assumed zero. Similarly if c_{12}, c_{13}, c_{23} were set equal to zero we should have the possibility of three extensional modes and one shear mode vibrating simultaneously with no reaction on one another, and the equation of motion would be

$$\begin{aligned} \rho \frac{\partial^2 u}{\partial t^2} &= \frac{\partial}{\partial x}(c_{11}x_x), \\ \rho \frac{\partial^2 v}{\partial t^2} &= \frac{\partial}{\partial y}(c_{22}y_y) + \frac{\partial}{\partial_z}[c_{44}y_z], \\ \rho \frac{\partial^2 w}{\partial t^2} &= \frac{\partial}{\partial y}(c_{44}y_z) + \frac{\partial}{\partial z}[c_{33}z_z]. \end{aligned} \tag{29}$$

The displacements u, v, and w would be the sum of the displacements caused by the four motions. To find the displacements and resonances caused by the shear mode y_z, we neglect the other modes and have the equations

$$\begin{aligned} \rho \frac{\partial^2 v}{\partial t^2} &= c_{44} \frac{\partial}{\partial z}(y_z), \\ \rho \frac{\partial^2 w}{\partial t^2} &= c_{44} \frac{\partial}{\partial y}(y_z). \end{aligned} \tag{30}$$

Differentiating the first of equations (30) by $\frac{\partial}{\partial z}$, and the second by $\frac{\partial}{\partial y}$, and adding, there results,

$$\rho \frac{\partial^2}{\partial t^2}\left(\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}\right) = c_{44}\left[\frac{\partial^2 y_z}{\partial y^2} + \frac{\partial^2 y_z}{\partial z^2}\right]. \tag{31}$$

[19] For example if motion is limited to the y and z directions, and the coefficient c_{14} is set equal to zero, the equations reduce approximately to those for a plate bent in flexure, and this case has never been solved for the boundary condition of interest here, namely all four edges being free to move—see Rayleigh "Theory of Sound," page 372, Vol. I, 1923 edition.

Since $\frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} = y_x$, this reduces to

$$\frac{\partial^2 y_x}{\partial t^2} = c^2 \left[\frac{\partial^2 y_x}{\partial y^2} + \frac{\partial^2 y_x}{\partial z^2} \right], \tag{32}$$

where $c^2 = c_{44}/\rho$.

For a simple harmonic vibration, of frequency f, the equation reduces to

$$\left[\frac{\partial^2 y_x}{\partial y^2} + \frac{\partial^2 y_x}{\partial z^2} \right] + \frac{p^2}{c^2} y_x = 0, \tag{33}$$

where $p = 2\pi f$. The solution of this equation consistent with the boundary conditions (28) is

$$y_x = \Sigma\Sigma A_{mn} \left[\sin \frac{m\pi y}{a} \sin \frac{n\pi z}{b} \right] \cos pt, \tag{34}$$

where a is the length of the crystal in the y direction, b the length of the crystal in the z direction, and m and n are integers. Substituting this equation in the equation (32), we find that it is a solution provided

$$\left[\frac{m^2\pi^2}{a^2} + \frac{n^2\pi^2}{b^2} \right] = \frac{p^2}{c^2} = \frac{(2\pi f)^2}{c^2}. \tag{35}$$

Hence the resonant frequencies of the crystal in shear vibration are

$$f = \frac{c}{2} \sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}. \tag{36}$$

To find the shape of the deformed crystal, we have from (30) for simple harmonic vibration

$$v = -\frac{c^2}{p^2} \frac{\partial y_x}{\partial z} = \frac{-a^2b^2}{m^2\pi^2b^2 + n^2\pi^2a^2} \frac{\partial y_x}{\partial z}, \tag{37}$$

$$w = -\frac{c^2}{p^2} \frac{\partial y_x}{\partial y} = \frac{-a^2b^2}{m^2\pi^2b^2 + n^2\pi^2a^2} \frac{\partial y_x}{\partial y}. \tag{38}$$

The cases $m = 0$, $n = 1$ and $m = 1$, $n = 0$ require a stress known as a simple shear to excite them, whereas the stress applied by the piezo-electric effect is a pure shear. Hence the case $m = 1$, $n = 1$ provides the lowest frequency solution. The displacements v and w for this case are by equations (34), (37) and (38)

$$\begin{aligned} v &= \frac{-a^2b\pi}{\pi^2a^2 + \pi^2b^2} \left(\sin \frac{\pi y}{a} \cos \frac{\pi z}{b} \right); \\ w &= -\frac{ab^2\pi}{\pi^2a^2 + \pi^2b^2} \left(\cos \frac{\pi y}{a} \sin \frac{\pi z}{b} \right). \end{aligned} \tag{39}$$

The resulting distortion of the crystal is shown by Fig. 34.

We can conclude, therefore, that the solution for a shear vibration in a quartz crystal will be given by equation (34). It is obvious from Fig. 34 that the shear vibration will have a strong coupling to a bar

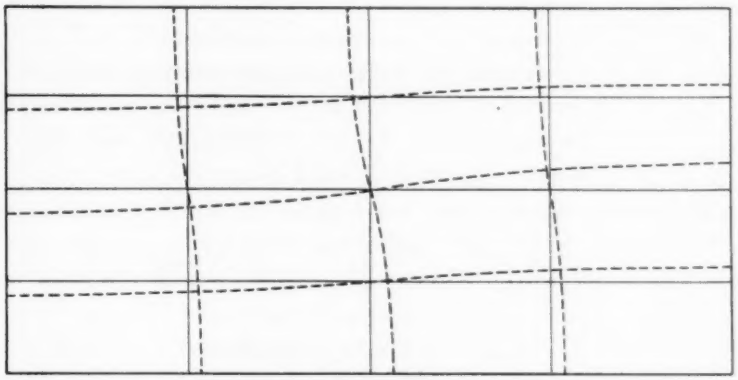


Fig. 34—Form of crystal in shear vibration.

bent in its second mode of flexure, since the form of the bent bar, as shown by Fig. 29, is very closely the same as a given displacement line in the crystal vibrating in shear. Little coupling should exist between the shear mode and a bar in its first flexure mode, since this mode of flexure requires a displacement which is symmetrical on both sides of the central line whereas the bar vibrating in shear has a motion in which the displacement on one side of the center line is the opposite of the displacement on the other side of the center line.

The Elastic and Piezo-Electric Constants of Quartz for Rotated Crystals [20]

W. Voigt [21] gives for the stress strain and piezo-electric relation in a quartz crystal, for the three extensions and one shear found above,

$$\begin{aligned} -x_x &= s_{11}X_x + s_{12}Y_y + s_{13}Z_z + s_{14}Y_z, \\ -y_y &= s_{12}X_x + s_{22}Y_y + s_{23}Z_z + s_{24}Y_z, \\ -z_z &= s_{13}X_x + s_{23}Y_y + s_{33}Z_z + s_{34}Y_z, \\ -y_z &= s_{14}X_x + s_{24}Y_y + s_{34}Z_z + s_{44}Y_z, \end{aligned} \tag{40}$$

$$-P_x = d_{11}X_x + d_{12}Y_y + d_{13}Z_z + d_{14}Y_z, \tag{41}$$

[20] The material of this section was first derived by Mr. R. A. Sykes of the Bell Telephone Laboratories.

[21] W. Voigt, Lehrbach Der Kristallphysik.

where

x_x, y_y, z_z = extensional strains = elongation per unit length,
X_x, Y_y, Z_z = extensional stresses = force per unit area,
y_z = shearing strain = cos of an angle,
Y_z = shearing stress = force per unit area,
s_{ij} = elastic compliances = displacement per dyne,
d_{ij} = piezo-electric constants = e.s.u. charge per dyne,
P_x = piezo-electric polarization = charge per unit area.

The best measured values for these constants when the X axis coincides with the electric axis of the crystal, the Y axis with the mechanical axis and the Z axis with the optical axis, are

$$\begin{aligned}
s_{11} &= s_{22} = 127.2 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
s_{12} &= -16.6 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
s_{13} &= s_{23} = -15.2 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
s_{24} &= -s_{14} = 43.1 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
s_{33} &= 97.2 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
s_{34} &= 0, \\
s_{44} &= 200.5 \times 10^{-14} \text{ cm.}^2/\text{dyne}, \\
d_{11} &= -d_{12} = -6.36 \times 10^{-8} \frac{\text{e.s.u. charge}}{\text{dyne}}, \\
d_{13} &= 0 \\
d_{14} &= 1.69 \times 10^{-8} \frac{\text{e.s.u. charge}}{\text{dyne}}.
\end{aligned} \tag{42}$$

If, now, we maintain the direction of the electrical axis but rotate the direction of the principal axis by some angle θ, the resulting constants of equations (40) and (41) undergo a change.

Let the direction cosines for the new axes be given by

	x	y	z
x'	l_1	m_1	n_1
y'	l_2	m_2	n_2
z'	l_3	m_3	n_3

(43)

The convention is adopted that a positive angle θ is a clockwise rotation of the principal axis of the crystal, when the electrically positive face (determined by a squeeze) is up. For a left-handed crystal a positive angle is in a counter clockwise direction. θ is the angle between the previously unprimed and the primed axes.

If we transform only the y and z axes, there results

$$\begin{aligned} l_2 &= l_3 = m_1 = n_1 = 0, \\ l_1 &= 1, \\ m_2 &= n_3 = \cos\theta, \\ -n_2 &= m_3 = \sin\theta. \end{aligned} \tag{44}$$

Love [22] gives the transformation for the stress and the strain functions as

$$\begin{aligned} x_x' &= x_x l_1^2 + y_y m_1^2 + z_z n_1^2 + y_z m_1 n_1, \\ y_y' &= x_x l_2^2 + y_y m_2^2 + z_z n_2^2 + y_z m_2 n_2, \\ z_z' &= x_x l_3^2 + y_y m_3^2 + z_z n_3^2 + y_z m_3 n_3, \\ y_z' &= 2x_x l_2 l_3 + 2y_y m_2 m_3 + 2z_z n_2 n_3 + y_z(m_2 n_3 + m_3 n_2), \end{aligned} \tag{45}$$

$$\begin{aligned} X_x' &= X_x l_1^2 + Y_y m_1^2 + Z_z n_1^2 + Y_z 2m_1 n_1, \\ Y_y' &= X_x l_2^2 + Y_y m_2^2 + Z_z n_2^2 + Y_z 2m_2 n_2, \\ Z_z' &= X_x l_3^2 + Y_y m_3^2 + Z_z n_3^2 + Y_z 2m_3 n_3, \\ Y_z' &= X_x l_2 l_3 + Y_y m_2 m_3 + Z_z n_2 n_3 + Y_z(m_2 n_3 + m_3 n_2). \end{aligned} \tag{46}$$

Substituting (44) in (45) and (46) and then expressing x_x, y_y . . . X_x, Y_y . . ., etc., in terms of x_x', y_y' . . . X_x', Y_y' . . ., etc., we may substitute these values in equations (40) and (41) to give the stress-strain and polarization in a crystal whose rectangular axes do not coincide with the real optical and mechanical axis. Performing the above operations, a new set of constants s_{ij}', are obtained which are functions of θ, namely:

$$s_{11}' = s_{11},$$

$$s_{12}' = \tfrac{1}{2}[s_{12} + s_{13} + (s_{12} - s_{13})\cos 2\theta - s_{14}\sin 2\theta],$$

$$s_{13}' = \tfrac{1}{2}[s_{13} + s_{12} + (s_{13} - s_{12})\cos 2\theta + s_{14}\sin 2\theta],$$

$$s_{14}' = s_{14}\cos 2\theta + (s_{12} - s_{13})\sin 2\theta,$$

$$s_{22}' = s_{11}\cos^4\theta + s_{33}\sin^4\theta + 2s_{14}\cos^3\theta\sin\theta + (2s_{13} + s_{44})\sin^2\theta\cos^2\theta,$$

$$s_{23}' = s_{13}(\cos^4\theta + \sin^4\theta) + s_{14}(\sin^3\theta\cos\theta - \cos^3\theta\sin\theta) + (s_{11} + s_{33} - s_{44})\sin^2\theta\cos^2\theta,$$

$$s_{24}' = -s_{14}(\cos^4\theta - 3\sin^2\theta\cos^2\theta) + (2s_{11} - 2s_{13} - s_{44})\cos^3\theta\sin\theta + (2s_{13} - 2s_{33} + s_{44})\sin^3\theta\cos\theta,$$

$$s_{33}' = s_{33}\cos^4\theta + s_{11}\sin^4\theta - 2s_{14}\sin^3\theta\cos\theta + (2s_{13} + s_{44})\sin^2\theta\cos^2\theta,$$

[22] "The Mathematical Theory of Elasticity," Cambridge University Press, pp. 42 and 78.

$$s_{34}' = s_{14}(\sin^4\theta - 3\sin^2\theta\cos^2\theta) + (2s_{11} - 2s_{13} - s_{44})\sin^3\theta\cos\theta + (2s_{13} - 2s_{33} + s_{44})\cos^3\theta\sin\theta,$$

$$s_{44}' = (4s_{33} + 4s_{11} - 8s_{13} - 2s_{44})\sin^2\theta\cos^2\theta + 4s_{14} \times (\sin^3\theta\cos\theta - \sin\theta\cos^3\theta) + s_{44}(\sin^4\theta + \cos^4\theta)$$

and

$$d_{11}' = d_{11},$$
$$d_{12}' = -\tfrac{1}{2}[d_{11}(1 + \cos 2\theta) + d_{14}\sin 2\theta],$$
$$d_{13}' = -\tfrac{1}{2}[d_{11}(1 - \cos 2\theta) - d_{14}\sin 2\theta],$$
$$d_{14}' = d_{14}\cos 2\theta - d_{11}\sin 2\theta.$$

The curves representing the s' values for varying angles of orientation are plotted on Fig. 30 while the values of d' are plotted on Fig. 35.

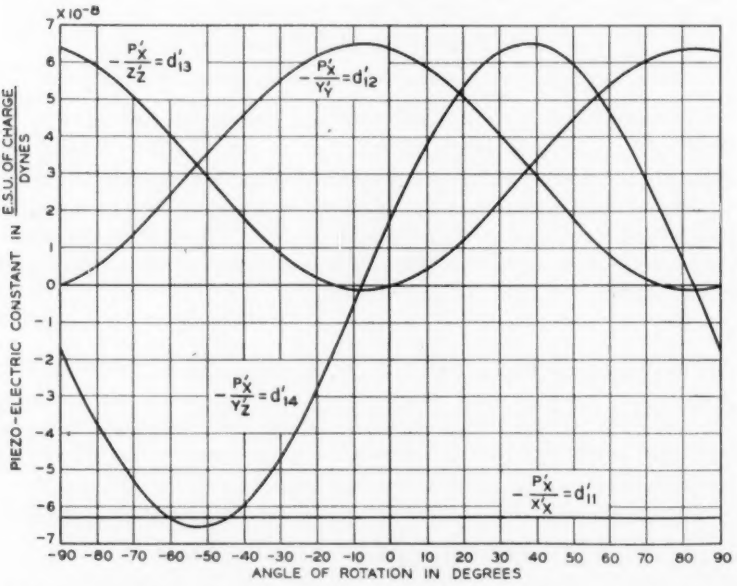


Fig. 35—Piezo-electric constants of a perpendicularly cut crystal as a function of the angle of rotation.

Some Improvements in Quartz Crystal Circuit Elements

By F. R. LACK, G. W. WILLARD, and I. E. FAIR

The characteristics of the *Y*-cut quartz crystal plate are discussed. It is shown that by rotating a plate about the *X* axis special orientations are found for which the frequency spectrum is simplified, the temperature coefficient of frequency is reduced practically to zero and the amount of power that can be controlled without fracture of the crystal is increased. These improvements are obtained without sacrificing the advantages of the Y cut plate, i.e., activity and the possibility of rigid clamping in the holder.

THERE are at the present time two types of crystal quartz plates in general use as circuit elements for frequency stabilization at radio frequencies, namely, the *X*-cut and *Y*-cut.[1] This paper is concerned with the improved characteristics of plates having radically new orientations.

In its usual form the *Y*-cut plate is cut from the mother crystal, as shown in Fig. 1. The electric field is applied along the *Y* direction

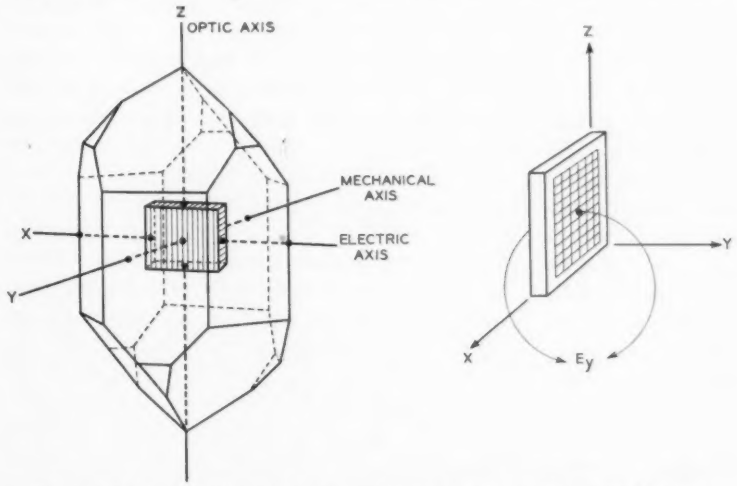


Fig. 1—Showing relation of Y-cut quartz crystal to the crystallographic axes.

and for high frequencies an x_y shear vibration is utilized.

The frequency of such a vibration is given by the expression:

$$f = \frac{1}{2l}\sqrt{\frac{c_{66}}{\rho}}, \qquad (1)$$

[1] "Piezo-Electric Terminology," W. G. Cady, *Proc. I. R. E.*, 1930, p. 2136.

where c_{66} = the elastic constant for quartz connecting the X_y stress with an x_y strain = 39.1×10^{10} dynes per cm.[2]

ρ = the density of quartz = 2.65 gms. per cm.[3]

l = the thickness in cm.

On substituting the numerical values in equation (1), a frequency-thickness constant of 192 kc. per cm. is obtained which checks within 3 per cent the value of this constant found by experiment.

This x_y shear vibration is not appreciably affected when the plate is rigidly clamped, the clamping being applied either around the periphery if the plate is circular, or at the corners if square. Hence a mechanically rigid holder arrangement is possible which is particularly suitable for mobile radio applications.[2]

The temperature coefficient of frequency of this vibration is approximately + 85 parts/million/C.°, which means that for most applications it must be used in a thermostatically controlled oven. In operation, this comparatively large temperature coefficient is responsible for a major part of any frequency deviations from the assigned value.

Another important characteristic of the *Y*-cut crystal is the secondary frequency spectrum of the plate. This secondary spectrum consists of overtones of low frequency vibrations which are mechanically coupled to the desired vibration and cause discontinuities in the characteristic frequency-temperature and frequency-thickness curves of the crystal. In some instances these coupled secondary vibrations can be utilized to produce a low temperature coefficient over a limited temperature range.[3] But in general, at the higher frequencies (above one megacycle) this secondary spectrum is a source of considerable annoyance, not only in the initial preparation of a plate for a given frequency but in the added necessity for some form of temperature control. In practice, these plates are so adjusted that there are no discontinuities in the frequency-temperature characteristic in the region where they are expected to operate, but at high frequencies it is difficult to eliminate all of these discontinuities over a wide temperature range. If, then, for any reason the crystal must be operated without the temperature control, a frequency discontinuity with temperature may cause a large frequency shift greatly in excess of that to be expected from the normal temperature coefficient.

From the above considerations it may be concluded that the standard *Y*-cut plate has two distinct disadvantages: namely, a

[2] U. S. Patent No. 1883111, G. M. Thurston, Oct. 18, 1932. "Application of Quartz Plates to Radio Transmitters," O. M. Hovgaard, *Proc. I. R. E.*, 1932, p. 767.

[3] "Observations on Modes of Vibrations and Temperature Coefficients of Quartz Plates," F. R. Lack, *Proc. I. R. E.*, 1929, p. 1123; *Bell Sys. Tech. Jour.*, July, 1929.

temperature coefficient requiring close temperature regulation and a troublesome secondary frequency spectrum. Assuming that the temperature coefficient of the desired frequency could be materially reduced, the effect of any secondary spectrum must also be minimized before temperature regulation can be abandoned. In fact, from the standpoint of satisfactory production and operation of these crystal plates, it is perhaps more important that the secondary spectrum be eliminated than that the temperature coefficient be reduced.

The Secondary Spectrum

The secondary spectrum of these plates, as has been indicated above, is caused by vibrations of the same or of other types than the wanted vibration taking place in other directions of the plate and coupled to the wanted vibration mechanically. This condition of affairs exists in all mechanical vibrating systems but is complicated in the case of quartz by the complex nature of the elastic system involved.

Considering specifically the case of the Y-cut plate the desired vibration is set up through the medium of an x_y strain. Hence any coupled secondary vibrations must be set in motion through this x_y strain. Referring to the following elastic equations for quartz (in these equations X, Y and Z are directions coincident with the crystallographic axes; see Fig. 1 and Appendix),

$$\left.\begin{aligned}
-X_x &= c_{11}x_x + c_{12}y_y + c_{13}z_z + c_{14}y_z \\
-Y_y &= c_{12}x_x + c_{22}y_y + c_{23}z_z + c_{24}y_z \\
-Z_z &= c_{13}x_x + c_{23}y_y + c_{33}z_z \\
-Y_z &= c_{14}x_x + c_{24}y_y + c_{44}y_z \\
-Z_x &= + c_{55}z_x + c_{56}x_y \\
-X_y &= + c_{56}z_x + c_{66}x_y
\end{aligned}\right\} \qquad (2)$$

it will be seen that by reason of the constant c_{56} an x_y strain will set up a stress in the Z_x plane which in turn will produce a z_x strain. Hence the x_y and z_x strains are coupled together mechanically, the value of the constant c_{56} being a measure of that coupling.

High order overtones of vibrations resulting from this z_x strain constitute the major part of the secondary frequency spectrum of these plates.

The technique for dealing with this secondary spectrum in the past has been the proper choice of dimensions. At high frequencies these overtones occur very close together and when one set is moved out of the range by grinding a given dimension another set will appear. Some benefit is obtained with the clamped holder, which tends to inhibit certain types of transverse vibrations; but as indicated above,

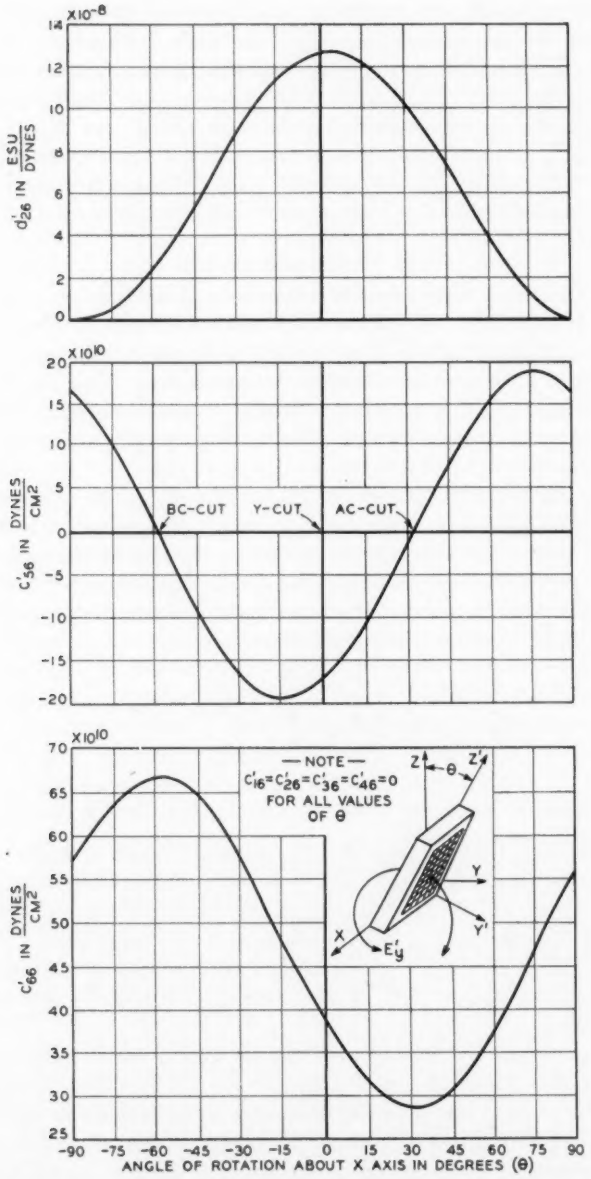


Fig. 2—c_{66}', c_{56}', and d_{26}' as a function of rotation about the X axis.

the elimination of the effects of the coupled secondary frequency spectrum over a wide temperature range is a difficult matter.

Another method has been developed recently for dealing with these coupled vibrations.[4] This consists of reducing, by a change in orientation, the magnitude of the elastic constant responsible for the coupling. If the orientation of the crystal plate be shifted with respect to the crystallographic axes then, in general, the elastic constants with reference to the axes of the plate will vary. The direct constants (c_{11}', c_{22}', $\cdots$) which represent the longitudinal and shear moduli will of course vary in magnitude only, while the cross constants (c_{ij}', $\cdots$) will vary both as to magnitude and sign. There is a possibility therefore that the proper choice of orientation of the plate will reduce c_{56}' to zero without at the same time introducing other couplings.

Figure 2 shows graphically the variation of c_{66}' and c_{56}' as a function of rotation about the X axis. These have been calculated by means of the equations given in the appendix. It will be seen that at approximately $+ 31°$ and $- 60°$ c_{56}' becomes zero. Here then are two orientations for which the coupling between the x_y and z_x strains should be zero.

In shifting the orientation of the plate the necessity for exciting the wanted vibration piezo-electrically must not be lost sight of. Hence in addition to computing the values of the elastic constants the variation of the piezo-electric moduli as a function of orientation must also be examined. Figure 2 also shows the effect of rotation about the X axis on d_{26}' (the constant connecting the E_y' electric field with the x_y' strain). It will be seen that at both $+ 31°$ and $- 60°$ the x_y' vibration can be excited piezo-electrically but it is to be expected that a plate cut at $- 60°$ will be relatively inactive,[5] for d_{26}' at this point is only 20 per cent of its value for the Y-cut plate. On the other hand at $+ 31°$ a plate would be practically equivalent to the Y-cut as far as activity is concerned. The frequency of the x_y' vibration for these special orientations can be calculated by means of equation (1) substituting for c_{66} the value of c_{66}' for the given angle as read from the curve of Fig. 2.

[4] The expression of the coupling between two modes of vibration in quartz in terms of the elastic constants was first suggested in 1930 by Mr. W. P. Mason of the Bell Telephone Laboratories.

[5] The word "activity" is a rather loose term used by experimenters in this field to describe the ease with which a given vibration can be excited in a particular circuit. It is often spoken of in terms of the grid current that is obtained in that circuit or the amount of feedback necessary to produce oscillation. It can better be expressed quantitatively as the coupling between the electrical and mechanical systems (not to be confused with the mechanical coupling between different vibrations described above) which is a simple function of the piezo-electric and elastic moduli of the vibration involved and the dielectric constant of the crystal plate.

For the purpose of identification the plate cut at + 31° has been designated as the *AC*-cut and the plate cut at − 60° the *BC*-cut. Crystal plates having these orientations have been made up and tested. It is evident from the frequency-temperature and frequency-thickness characteristics of both cuts that a simplification of the frequency spectrum results from the reduction in coupling to secondary

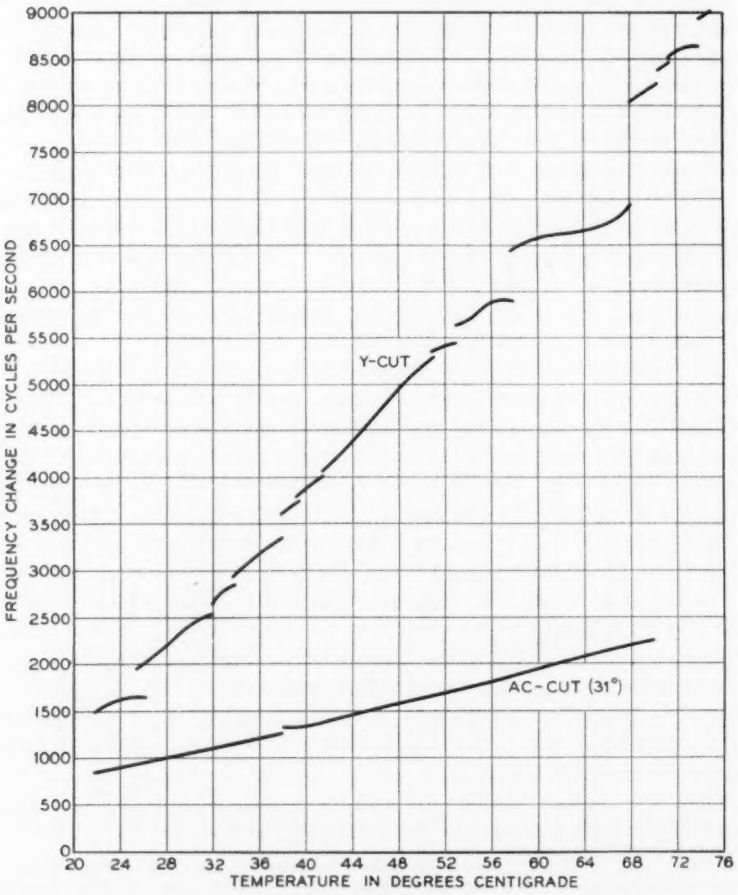


Fig. 3—Frequency-temperature characteristics of AC-cut and Y-cut plates of same frequency and area.

Frequency 1600 KC.
Dimensions:

Y-cut	y = 1.22 mm.	x = 38 mm.	z = 38 mm.
AC-cut (31°)	y' = 1.00 mm.	x = 38 mm.	z' = 38 mm.

vibrations. Frequency discontinuities of the order of a kilocycle or more which are a common occurrence with the Y-cut plate have disappeared and frequency-temperature curves that are linear over a considerable temperature range can be obtained without much difficulty. This is illustrated by Fig. 3 which shows frequency-temperature characteristics for both AC-cut and Y-cut plates of the same frequency and area. The AC-cut plate can be clamped to the same extent as the Y-cut plate.

There is still some coupling remaining to certain secondary frequencies. These frequencies are difficult to identify but are thought to be caused by overtones of flexural vibrations set up by the x_y' shear itself and hence would be unaffected by the reduction of c_{56}'. These remaining frequencies do not cause much difficulty above 500 kc. For the AC-cut ($+ 31°$) plate the temperature coefficient of frequency is $+ 20$ cycles/million/C.°, while for the BC-cut ($- 60°$) plate it is $- 20$ cycles/million/C.°.

In addition to these calculations for the x_y' vibration in plates rotated about the X axis, a detailed study has been made of other types of vibration and rotation about the other axes. For high frequencies nothing has been found to compare with the reduction in complexity of frequency spectrum obtainable with these two orientations.

Temperature Coefficients

This study has produced in the AC-cut a new type of plate which has superior characteristics to the standard Y-cut: i.e., a simplified frequency spectrum and a lower temperature coefficient. The values of the temperature coefficients obtained for these new orientations are significant and suggest that perhaps other orientations can be chosen for which the temperature coefficient will be zero. With the measured values of the temperature coefficients for the different orientations and the c_{66}' equation (Appendix) it is possible to compute the temperature coefficient for any angle. Figure 4 shows graphically the results of such a computation for an x_y' vibration as a function of rotation about the X axis. It will be seen that at approximately $+ 35°$ and $- 49°$ the x_y' vibration will have a zero temperature coefficient of frequency.

This curve has been checked experimentally, the check points being indicated on the curve. Concentrating on a plate cut at $+ 35°$, which has been designated the AT-cut, it will be seen that this type of plate offers considerable possibilities. Figure 5 shows the frequency-temperature curves of a 2-megacycle AT-cut plate and a

standard Y-cut of the same frequency and area. These curves not only illustrate the reduction in temperature coefficient but also show that in the AT-cut plate the secondary frequency spectrum has been eliminated over the temperature range of the test. This is to be expected, for 35° is close to the 31° zero coupling point; hence such coupling as does exist is small in magnitude.

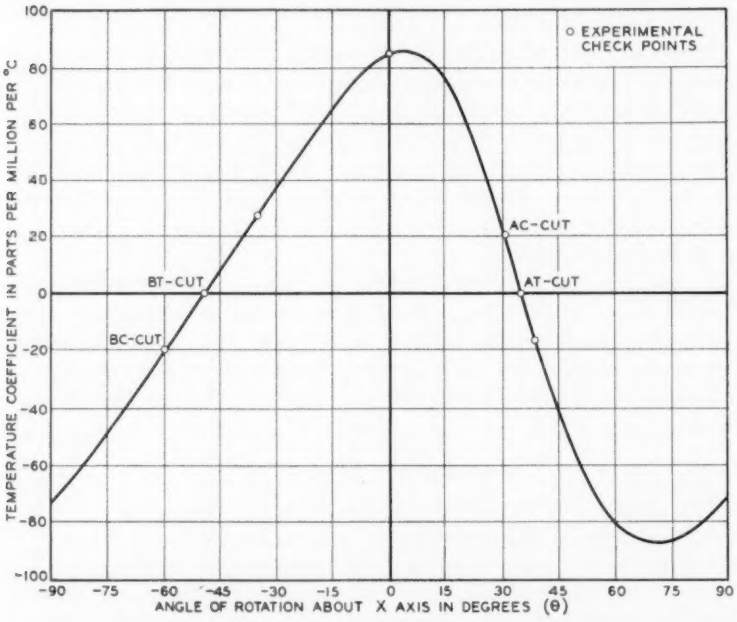


Fig. 4—Temperature coefficient of frequency of the vibration depending upon c_{66}' as a function of rotation about the X axis.

These AT-cut plates can be produced with a sufficiently low temperature coefficient so that for most applications the temperature regulating system can be discarded, and in addition a simplification of the secondary frequency spectrum is obtained. Furthermore, the advantages of the Y-cut plate, i.e., clamping and activity, have not been sacrificed.

Additional tests on AT-cut plates indicate that it will be possible to use them to control reasonable amounts of power without danger of fracture. At 2 megacycles, 50-watt crystal oscillators would appear to be practical and in some experimental circuits the power output has been run up to 200 watts without fracturing the crystal. The

explanation for this lies in the fact that the reduction in magnitude of the coupling to transverse vibrations has reduced the transverse stresses which in the *Y*-cut plate are responsible for the fractures.

Experimental crystals of this type have been produced in the frequency range from 500 kc. to 20 megacycles. The possibility of high frequencies, together with the elimination of the temperature

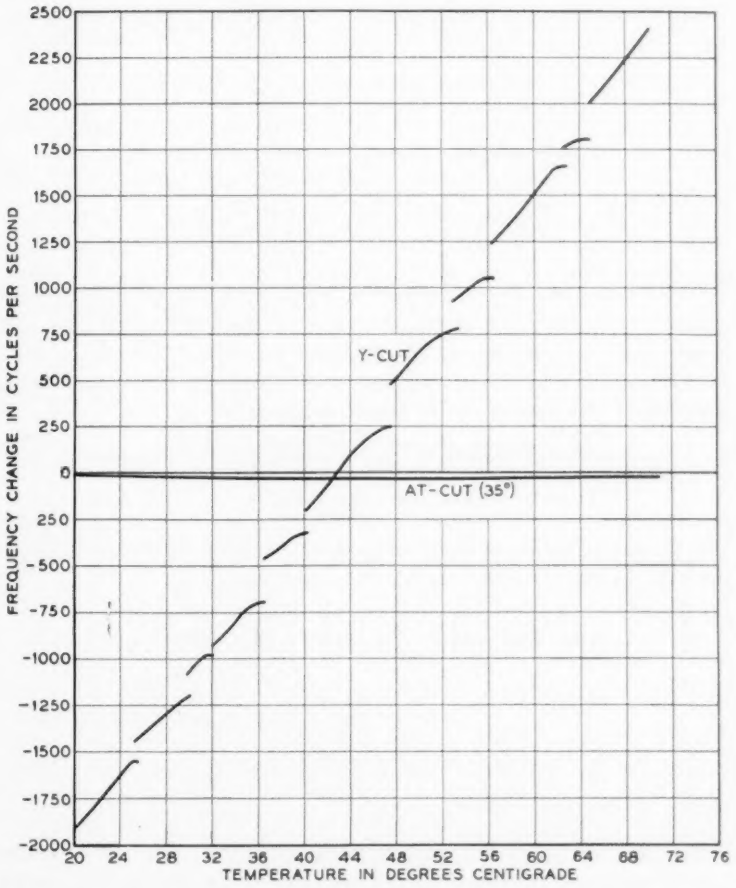


Fig. 5—Frequency-temperature characteristics of AT-cut and Y-cut plates of same frequency and area.

Frequency 1000 KC.
Dimensions:

Y-cut	y = 1.970 mm.	x = 38 mm.	z = 38 mm.
AT-cut (35°)	y' = 1.675 mm.	x = 38 mm.	z' = 38 mm.

control and the increase in the amount of power that may be controlled, should result in a considerable simplification of future short wave radio equipment.

APPENDIX

Elastic Equations

The general elastic equations for any crystal are given below, X', Y' and Z' representing any orthogonal set of axes.

$$\left.\begin{aligned}
-X_x' &= c_{11}'x_x' + c_{12}'y_y' + c_{13}'z_z' + c_{14}'y_z' + c_{15}'z_x' + c_{16}'x_y' \\
-Y_y' &= c_{12}'x_x' + c_{22}'y_y' + c_{23}'z_z' + c_{24}'y_z' + c_{25}'z_x' + c_{26}'x_y' \\
-Z_z' &= c_{13}'x_x' + c_{23}'y_y' + c_{33}'z_z' + c_{34}'y_z' + c_{35}'z_x' + c_{36}'x_y' \\
-Y_z' &= c_{14}'x_x' + c_{24}'y_y' + c_{34}'z_z' + c_{44}'y_z' + c_{45}'z_x' + c_{46}'x_y' \\
-Z_x' &= c_{15}'x_x' + c_{25}'y_y' + c_{35}'z_z' + c_{45}'y_z' + c_{55}'z_x' + c_{56}'x_y' \\
-X_y' &= c_{16}'x_x' + c_{26}'y_y' + c_{36}'z_z' + c_{46}'y_z' + c_{56}'z_x' + c_{66}'x_y'
\end{aligned}\right\} \quad (3)$$

When in quartz X', Y' and Z' coincide with the crystallographic axes of the material (X the electric axis, Y the mechanical axis, and Z the optic axis), equation (3) reduces to equation (2) of the text. In addition the following relations exist between the constants of equation (2) because of conditions of symmetry

$$c_{11} = c_{22}, \qquad c_{44} = c_{55}, \qquad c_{66} = (c_{11} - c_{12})/2, \qquad c_{13} = c_{23}$$
$$c_{14} = -c_{24} = c_{56}.$$

The numerical values of these constants have been determined experimentally by Voigt [6] and others.

$$\begin{aligned}
c_{11} &= 85.1 \times 10^{10}\,\frac{\text{dy.}}{\text{cm.}^2} & c_{12} &= 6.95 \times 10^{10}\,\frac{\text{dy.}}{\text{cm.}^2}, \\
c_{33} &= 105.3 \times 10^{10}\,\frac{\text{dy.}}{\text{cm.}^2} & c_{13} &= 14.1 \times 10^{10}\,\frac{\text{dy.}}{\text{cm.}^2}, \\
c_{44} &= 57.1 \times 10^{10}\,\frac{\text{dy.}}{\text{cm.}^2} & c_{14} &= 16.8 \times 10^{10}\,\frac{\text{dy.}}{\text{cm.}^2}, \\
c_{66} &= 39.1.
\end{aligned}$$

Using these constants it is possible to calculate the c_{ij}' for any orientation by means of transformation equations.[7] The expressions giving c_{16}', c_{26}', $\cdots$ c_{66}' (the constants relating to the x_y' strain) in terms of the c_{ij} for rotation about the X axis, are given below, θ being the

[6] W. Voigt, "Lehrbuch der Kristallphysik," 1928, p. 754.
[7] A. E. H. Love, "Mathematical Theory of Elasticity," 4th ed., p. 43.

angle between the Z' and Z axis (Fig. 2).

$$\begin{aligned} c_{16}' &= c_{26}' = c_{36}' = c_{46}' = 0, \\ c_{56}' &= c_{14}(\cos^2\theta - \sin^2\theta) + (c_{66} - c_{44})\sin\theta\cos\theta, \\ c_{66}' &= c_{44}\sin^2\theta + c_{66}\cos^2\theta - 2c_{14}\sin\theta\cos\theta. \end{aligned} \tag{4}$$

Piezo-Electric Equations

The inverse piezo-electric relations for the X', Y', Z' system of axes can be expressed by the following equations:

$$\left.\begin{aligned} x_x' &= d_{11}'E_x' + d_{21}'E_y' + d_{31}'E_z' \\ y_y' &= d_{12}'E_x' + d_{22}'E_y' + d_{32}'E_z' \\ z_z' &= d_{13}'E_x' + d_{23}'E_y' + d_{33}'E_z' \\ y_z' &= d_{14}'E_x' + d_{24}'E_y' + d_{34}'E_z' \\ z_x' &= d_{15}'E_x' + d_{25}'E_y' + d_{35}'E_z' \\ x_y' &= d_{16}'E_x' + d_{26}'E_y' + d_{36}'E_z' \end{aligned}\right\} \tag{5}$$

When in quartz X', Y', Z' coincide with the crystallographic axes, eq. 5 reduces to the following:

$$\left.\begin{aligned} x_x &= d_{11}E_x \\ y_y &= -d_{11}E_x \\ z_z &= 0 \\ y_z &= d_{14}E_x \\ z_x &= -d_{14}E_y \\ x_y &= -2d_{11}E_y \end{aligned}\right\} \tag{6}$$

where

$$d_{11} = \mp 6.36 \times 10^{-8}\,\frac{\text{esu}}{\text{dyne}},$$

$$d_{14} = 1.69 \times 10^{-8}\,\frac{\text{esu}}{\text{dyne}}.$$

For rotation about the X axis,

$$d_{26}' = (d_{14}\sin\theta - 2d_{11}\cos\theta)\cos\theta. \tag{7}$$

A Theory of Scanning and Its Relation to the Characteristics of the Transmitted Signal in Telephotography and Television

By PIERRE MERTZ and FRANK GRAY

By the use of a two-dimensional Fourier analysis of the transmitted picture a theory of scanning is developed and the scanning system related to the signal used for the transmission. On the basis of this theory a number of conclusions can be drawn:

1. The result of the complete process of transmission may be divided into two parts, (*a*) a reproduction of the original picture with a blurring similar to that caused in general by an optical system of only finite perfection, and (*b*) the superposition on it of an extraneous pattern not present in the original, but which is a function of both the original and the scanning system.

2. Roughly half the frequency range occupied by the transmitted signal is idle. Its frequency spectrum consists of alternating strong bands and regions of weak energy. In the latter the signal energy reproducing the original is at its weakest, and gives rise to the strongest part of the extraneous pattern. In a television system these idle regions are several hundred to several thousand cycles wide and have actually been used experimentally as the transmission path for independent signaling channels, without any visible effect on the received picture.

3. With respect to the blurring of the original all reasonable shapes of aperture give about the same result when of equivalent size. The sizes (along a given dimension) are determined as equivalent when the apertures have the same radius of gyration (about a perpendicular axis in the plane of the aperture).

4. With respect to extraneous patterns certain shapes of aperture are better than others, but all apertures can be made to suppress them at the expense of blurring. An aperture arrangement is presented which almost completely eliminates extraneous pattern while about doubling the blurring across the direction of scanning as compared with the usual square aperture. From this and other examples the degradation caused by the extraneous patterns is estimated.

IN the usual telephotographic or television systems the image field is scanned by moving a spot or elementary area along some recurring geometrical path over this field. In the more common arrangement this path consists simply of a series of successive parallel strips. Imagining the path developed or straightened out (or in the more common case, the strips joined end to end), this method of scanning is equivalent to transmitting the image in the form of a long narrow strip.

The theoretical treatment of such transmission has usually been developed by completely ignoring variations in brightness across the image strip, assuming the brightness to have a uniform distribution across this strip. This permits the image to be analyzed as an ordinary one-dimensional or single Fourier series (or integral) along the length of the strip; and the theory is then developed in terms of the

one-dimensional steady state Fourier components. Such a method of treatment naturally gives no information in regard to the reproduction or distortion of the detail in the original image across the direction of scanning, nor, as will appear below, does it give any detailed information in regard to the fine-structure distribution of energy over the frequency range occupied by the signal.

The need of a more detailed theoretical treatment originally arose in connection with studies of the reproduction of detail in telephotographic systems, especially in comparisons of distortion occurring along the direction of scanning with that across this direction. Later, this same need was strikingly shown by the discovery that a television signal leaves certain parts of the frequency range relatively empty of current components. Certain considerations indicated that a large part of the energy of a signal might be located in bands at multiples of the frequency of line scanning. Actual frequency analyses more than confirmed this suspicion. The energy was found to be so closely confined to such bands as to leave the regions between relatively empty of signal energy.

Such bands and intervening empty regions are illustrated by the examples of current-frequency curves in Fig. 1. These curves were taken with the various subjects as indicated, and the television current was generated by an apparatus scanning a field of view in 50 lines at a rate of about 940 lines per second. The energy is grouped in bands at multiples of 940 cycles and the regions between are substantially devoid of current components. In addition to the bands shown by the curves, it is known that similar bands occur up to about 18,000 cycles and that there is also a band of energy extending up from about 20 cycles.

Certain of the relatively empty frequency regions were also investigated by including a narrow band elimination filter in a television circuit. The filter eliminated a band about 250 cycles wide and was variable so that the band of elimination could be shifted along the frequency scale at will. By shifting the region of elimination along in this manner it was found that a band about 500 or 600 cycles wide could be removed from a television channel between any two of the current components without producing any detectable effect on the reproduced image.

At a later date a 1500-cycle current suitable for synchronization was introduced into a relatively empty frequency region, transmitted over the same channel with a television current, and filtered out—all without visibly affecting the image.

These results indicated quite clearly the need of a more complete

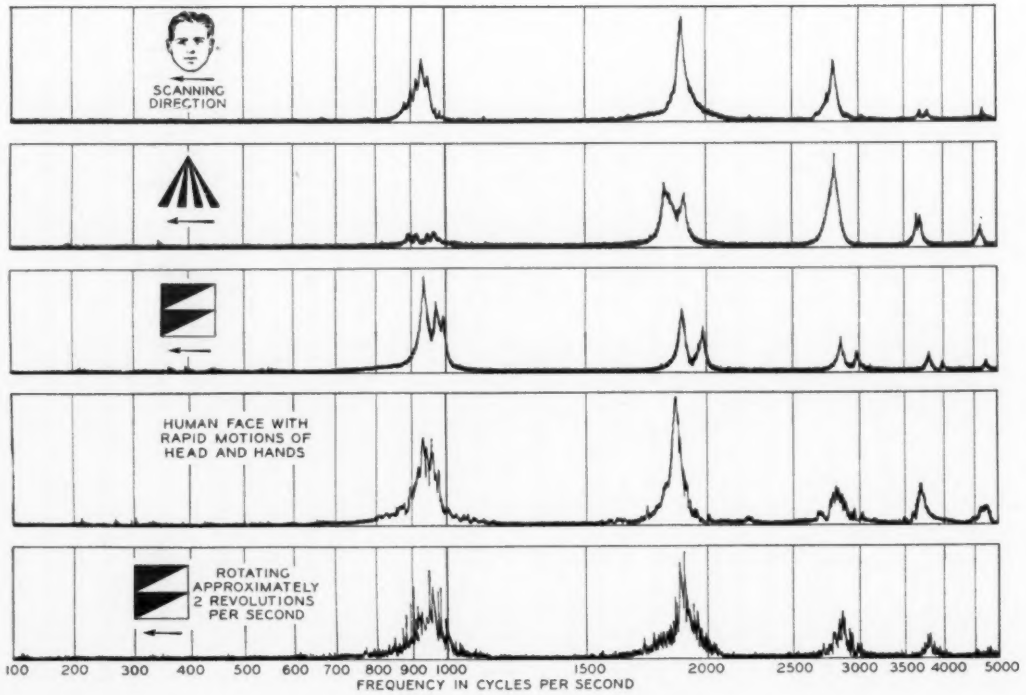


Fig. 1—Frequency analyses of television currents.

theory of the scanning processes used in telephotography and television and led to the study outlined in the following pages. Since this study will be confined to characteristics of the scanning processes all other processes in the system, wherever used, will be assumed to be perfect and cause no distortion.

The general trend of this more complete theory can be foreseen when it is considered that to obtain an adequate reproduction of the original it is necessary to scan with a large number of lines as compared with the general pictorial complexity of this original. This means that for any original presenting a large scale pattern (as distinguished from a random granular background) the signal pattern along successive scanning lines will, in general, differ by only small amounts. Thus, the signal wave throughout a considerable number of scanning lines may be represented to within a small error by a function periodic in the scanning frequency. Since such a function, developed in a Fourier series, is equal to the sum of sine waves having frequencies which are harmonics of the scanning line frequency, it will be natural to expect the total signal wave to have a large portion of its energy concentrated in the regions of these harmonics.

Furthermore, the existence of signal energy at odd multiples of half the scanning frequency will indicate the existence of a characteristic in the picture which repeats itself in alternate scanning lines. It is to be expected that such detail in a picture cannot be transmitted without accurate registry between it and the scanning lines and that when the detail spacing or direction or both differ somewhat from the scanning line spacing and direction, beat patterns between the two will be produced in the received picture which may be strong enough to alter considerably the reproduction of the original.

These phenomena are exactly what is observed, and will be treated in more quantitative fashion in the discussion below.[1]

An Image Field as a Double Fourier Series

Let us first consider the usual expression of the image field as a single Fourier series. The picture will be considered as a "still" so that entire successive scannings are identical. Then if the long strip corresponding to one scanning extends from $-L$ to $+L$, the illumina-

[1] In the following treatment an effort has been made to confine the necessary mathematical demonstrations almost exclusively to two sections entitled, respectively, "Effect of a Finite Aperture at the Transmitting Station," and "Reconstruction of the Image at the Receiving Station." Even in these sections a number of conclusions are explained in text which do not require reading the mathematics if the demonstrations are taken for granted. The occasional mathematical expressions occurring in the earlier sections are very largely for the purpose of introducing notation.

tion E as a function of the distance x along the strip may be expressed as the sum of an infinite number of Fourier components, thus:

$$E(x) = \sum_{n=0}^{\infty} a_n \cos\left(\frac{n\pi x}{L} + \varphi_n\right). \tag{1}$$

In this summation a_n represents the intensity of the nth component and φ_n its phase angle. The complete array of these for all components will vary if the picture is changed.

The cosine series above is very convenient for physical interpretation. It will be simple, however, for some of the later mathematical work to use the corresponding exponential series. The cosine series can be returned to, each time, as physical interpretation is required. That is, since

$$2a \cos\left(\frac{\pi x}{L} + \varphi\right) = (ae^{i\varphi})e^{(i\pi x/L)} + (ae^{-i\varphi})e^{(-i\pi x/L)} \tag{2}$$

the series in equation (1) can be written

$$E(x) = \sum_{n=-\infty}^{+\infty} A_n \exp i\pi(nx/L) \tag{3}$$

if we make

$$A_n = (1/2)a_n \exp(i\varphi_n)$$

and

$$A_{-n} = (1/2)a_n \exp(-i\varphi_n) \tag{4}$$

and if we use the notation $\exp\theta = e^{\theta}$.

In this new summation the complex amplitude A_n represents both the absolute intensity and the phase angle of the nth component. The complex amplitude of the corresponding component with a negative subscript is merely the conjugate of this.

As has already been noted, however, and as might readily be expected, the single Fourier series in equations (1) or (3) above do not always represent a two-dimensional picture with sufficient completeness. In order to consider the two-dimensional field more in detail, let us assume that Fig. 2 represents such an image field of dimensions $2a$ and $2b$, and take axes of reference x and y as indicated. The brightness or illumination of the field is a function $E(x, y)$ of both x and y. Along any horizontal line (i.e., in the x direction, constantly keeping $y = y_1$) the illumination may be expressed as a single Fourier series

$$E(x, y_1) = \sum_{m=-\infty}^{+\infty} A_m \exp i(mx/a). \tag{5}$$

Along any other line in the x direction a similar series holds with different coefficients, that is, the A's are functions of y. They may

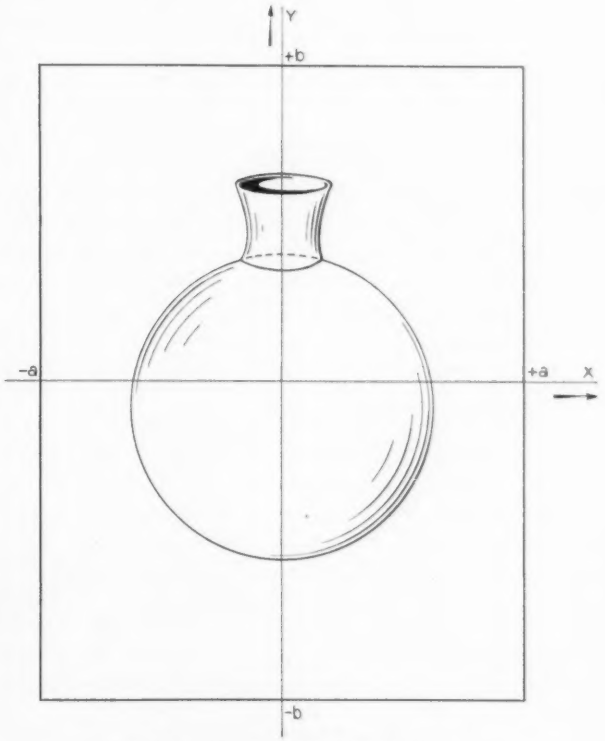


Fig. 2—Scanned field and image.

therefore each be written as a Fourier series along y

$$A_m = \sum_{n=-\infty}^{+\infty} A_{mn} \exp i\pi(ny/b). \tag{6}$$

Substitution in equation (5) gives the double Fourier series,

$$E(x, y) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} A_{mn} \exp i\pi \left(\frac{mx}{a} + \frac{ny}{b} \right). \tag{7}$$

For purposes of physical interpretation, as in the case of the simple Fourier series, it is desirable to combine the $+m$, $+n$ term with the $-m$, $-n$ term (giving the single $(m, +n)$th component) and similarly

the $+m, -n$ with the $-m, +n$ terms (giving the single $(m, -n)$th component). This brings equation (7) back to a cosine series,

$$E(x, y) = \sum_{m=0}^{\infty} \sum_{n=-\infty}^{+\infty} a_{mn} \cos \left[\pi \left(\frac{mx}{a} + \frac{ny}{b} \right) + \varphi_{mn} \right] \tag{8}$$

when

$$A_{mn} = (1/2) a_{mn} \exp (i\varphi_{mn})$$

and

$$A_{-m-n} = (1/2) a_{mn} \exp (-i\varphi_{mn})$$

and where a_{mn} is always a real quantity. Each term of this series represents a real, two-dimensional, sinusoidal variation in brightness extending across the image field. The image is built up of a superposition of a series of such waves extending across the field in various directions and having various wave lengths.

Imagining brightness as a third dimension, we may, as an aid in visualizing the components of an image field, draw separate examples of various components as shown in Fig. 3. It will be noted that any given component (m, n) passes through m periods along any horizontal line in the image field, and through n periods along any vertical line. The slope of the striations with respect to the x-axis is therefore $-mb/na$ (the negative reciprocal of the slope of the line of fastest variation in brightness). For the same values of m and of n, the $m, +n$ component and the $m, -n$ component have equal wave lengths but are sloped in opposite directions to the x-axis. If m is zero the crests are parallel to the x-axis; if n is zero they are parallel to the y-axis. The component with both m and n zero is a uniform distribution of brightness covering the entire image field. The wave length of a component is

$$\lambda_{mn} = 1 \Big/ \sqrt{\left(\frac{m}{2a} \right)^2 + \left(\frac{n}{2b} \right)^2}.$$

A complete array of the components, up to m and n equal to 4, is illustrated in Fig. 4.

As of course is characteristic of the harmonic analysis, the wave lengths and orientations of the components are seen to vary only with the shape and size of the rectangular field, and to be independent of the particular subject in the field. A change of subject, or motion of the subject, merely alters the amplitudes of the components and shifts their phase; but their wave length and inclination with respect to the x-axis remain unchanged. Consequently, for the same rectangular field all subjects appearing in it may be considered as built up from the same set of components. For a "still" subject, the amplitudes and

phase angles of the cosine components, or the complex amplitudes of the exponential components, remain constant with time. For a moving subject these complex amplitudes may be considered modulated as functions of time.

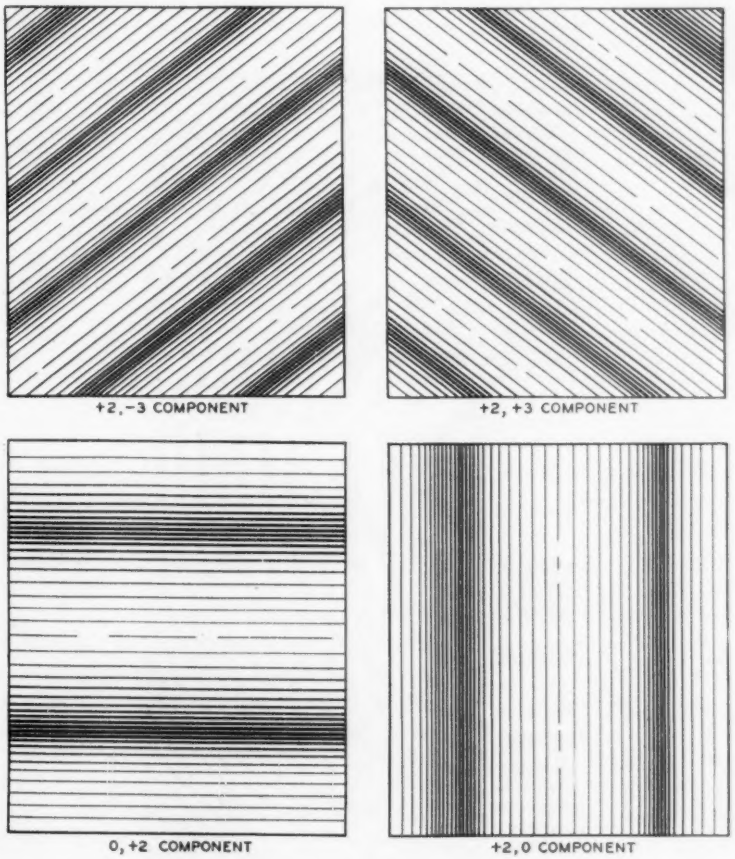


Fig. 3—Examples of field components.

The real amplitudes a_{mn} for a circular area of uniform brightness on a black background are relatively easily calculated, and this subject is also a good one to study as a picture from some points of view because it has a simple sharp border sloping in various directions. The amplitudes

for a circle of unit illumination of radius R are

$$a_{mn} = \frac{\pi R^2}{2ab}\left(\frac{\lambda_{mn}}{2\pi R}\right) J_1\left(\frac{2\pi R}{\lambda_{mn}}\right), \tag{9}$$

where J_1 is the first order Bessel function. In this particular subject all components of a given wave length have equal amplitudes; and the

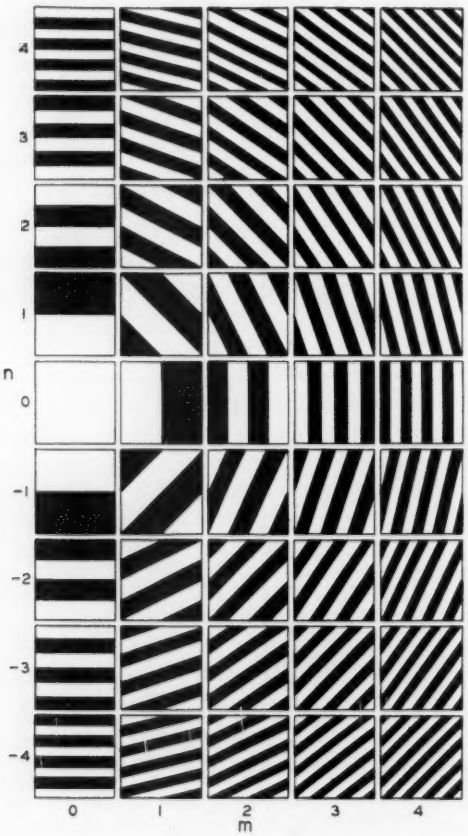


Fig. 4—Array of field components.

amplitudes may therefore be plotted as a function of wave length alone, as in Fig. 5. The curve illustrates the rapidity with which the amplitudes fall off for the higher order components in a subject of this nature.

The Frequency Spectrum of the Signal

When an image field is scanned by a point aperture tracing across it, each portion of the picture traversed causes variations in the light reaching the light sensitive cell and is thus translated into a corre-

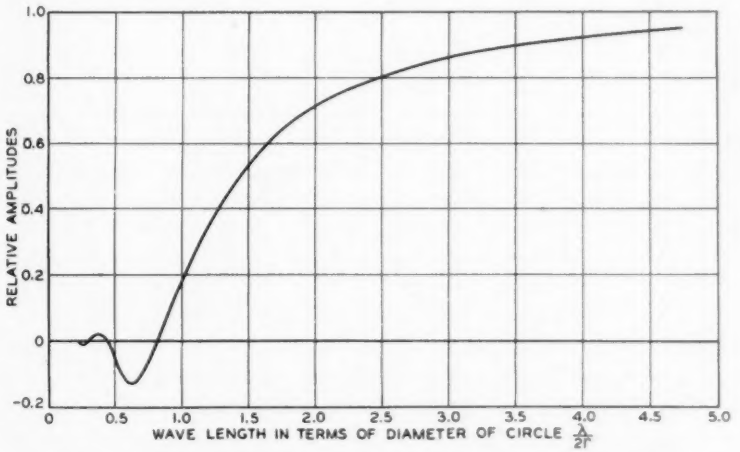


Fig. 5—Amplitudes of field components for a circular area of brightness.

sponding signal. Further, each Fourier component in the field is translated into a corresponding Fourier component in the signal. An equivalent translation occurs when a pencil of light traces over a photographic film in telephotography, or when a subject is scanned by a beam of light in television, whether or not a simple flat two-dimensional image is ever physically formed at the transmitting station. For clarity and simplicity, the discussion will be confined to the case in which a point aperture traces across a plane image field.

In most systems the aperture traces a line across the field and then there is a sudden jump back to the beginning of the next succeeding line. This discontinuous motion is naturally not easily subjected to mathematical treatment. It is much simpler to deal with the equivalent result that would be obtained if the scanning point, instead of tracing successive parallel paths across the same field, moved continuously across a series of identical fields. Such an equivalent scanning motion can fortunately easily be used because a double Fourier series represents not only a single field, but a whole succession of identical image fields covering the entire xy plane, and repeated periodically in both the x and y direction as illustrated in Fig. 6.

The equivalent of scanning a single field in parallel lines is obtained by assuming that the scanning point moves across the repeated fields along a sloping path as indicated. Let u be the velocity parallel to

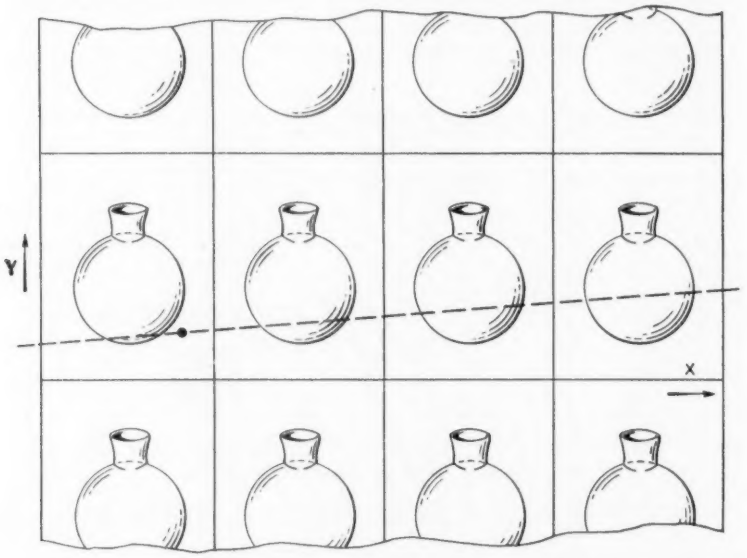


Fig. 6—Array of periodically recurring scanned fields.

the x axis and v the velocity parallel to the y axis. Then the picture illumination at the scanning point at any instant, and consequently the signal current, may be obtained by substituting

$$x = ut, \qquad y = vt$$

in the double Fourier series representing the image field, equation (8). Of course the entire expression must be multiplied by a factor K which is the constant ratio between the signal current and the picture illumination. This gives for the real signal as a function of time [2]

[2] It will be noted that this process does not explore the picture completely, inasmuch as, no matter how fine the scanning, there will always be unexplored regions between scanning lines. In this respect the process is quite analogous to that followed in analyzing a function of a single variable into a simple Fourier series when the values of the function are given only at discrete (even though closely spaced) values of the variable. The complete exact theory, which necessarily depends upon the size and shape of the finite scanning spot or aperture, will be given further below.

$$I(t) = K \sum_{m=0}^{\infty} \sum_{n=-\infty}^{+\infty} a_{mn} \cos \left[\pi \left(\frac{mu}{a} + \frac{nv}{b} \right) t + \varphi_{mn} \right]. \qquad (10)$$

Thus if u and v are constants, each wave of the image field gives rise to a corresponding Fourier component of the signal. The frequencies of the signal components are

$$f = \frac{mu}{2a} + \frac{nv}{2b}. \qquad (11)$$

The frequency spectrum of the signal is thus made up of a series of possible discrete lines, the position of which in that spectrum is determined by u and v, that is, by the particular scanning motion employed. We shall designate these lines by the indices m, $+n$ and m, $-n$, as they are correlated with the particular components of the image field that generated them.

A different choice of values for u and v (so long as these, once having been chosen, remain constant) changes the location of the lines in the frequency spectrum, but their amplitudes, depending only on the corresponding components of the image field, remain unchanged. In other words, the lines in the frequency spectrum of the signal are characteristic of the image field, and the scanning motion merely determines where they will appear in the frequency spectrum. Thus, if for a given subject the distribution of energy over the frequency scale is known for one method of scanning, it can be predicted for a great many other methods.

To scan a field in lines approximately parallel to the x axis, the velocity v must be made small compared to u. Under such conditions, $u/(2a)$ of equation (11) is the line scanning frequency and $v/(2b)$ is the frequency of image repetitions (or "frame frequency"). The frequency spectrum of the signal for a "still" picture thus consists of certain fundamental components at multiples of the line scanning frequency $u/(2a)$, each of which is accompanied by a series of lines spaced at equal successive intervals to either side of it. The spacing between these satellites is the image repetition or frame frequency $v/(2b)$.

If the picture changes with time the amplitudes of these fundamental lines and their satellites are modulated, also with respect to time. In other words they each develop sidebands or become diffuse. The diffuseness will not overlap from satellite to satellite unless the frequency of modulation becomes as great as half the frame frequency.

Thus for motions in the picture which are not too fast to be expected to be reproduced with reasonable fidelity, this diffuseness of the fundamental lines and their satellites will not obliterate their identity.

A diagrammatic arrangement of some of the possible lines in a frequency spectrum, with their corresponding m and n indices, is shown in Fig. 7.

It is important to note that the correlation between the wave lengths of the field components and the frequencies of the current components is not the one that is naturally assumed on first consideration. We

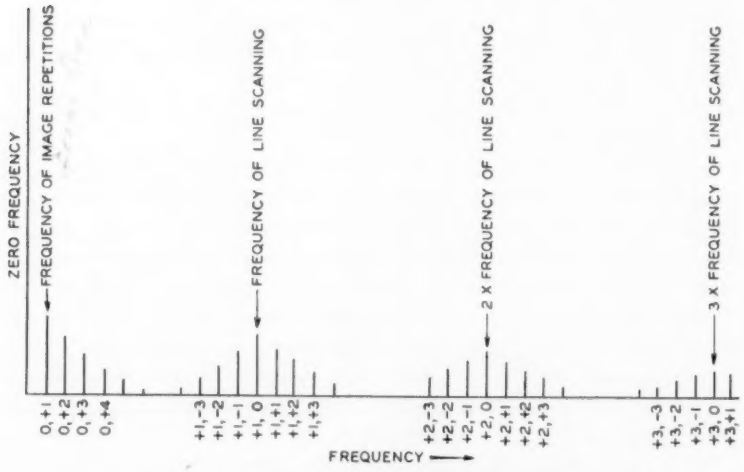


Fig. 7—Diagram of signal frequency spectrum.

are quite likely to make the erroneous assumption that high frequencies correspond to all sharp changes in brightness and that low frequencies correspond only to slow changes. The error in this assumption is readily realized by noting that sharp changes in brightness may generate very low frequencies if the scanning point passes over them in a sloping direction. An actual correlation is shown schematically in Fig. 8. It is seen that the same general type of correlation is repeated periodically over the frequency scale at multiples of the line scanning frequency. There are evidently numerous regions of the spectrum in which short image waves, or fine grained details of the image field, may appear in the signal. They are not confined to the high-frequency region alone.

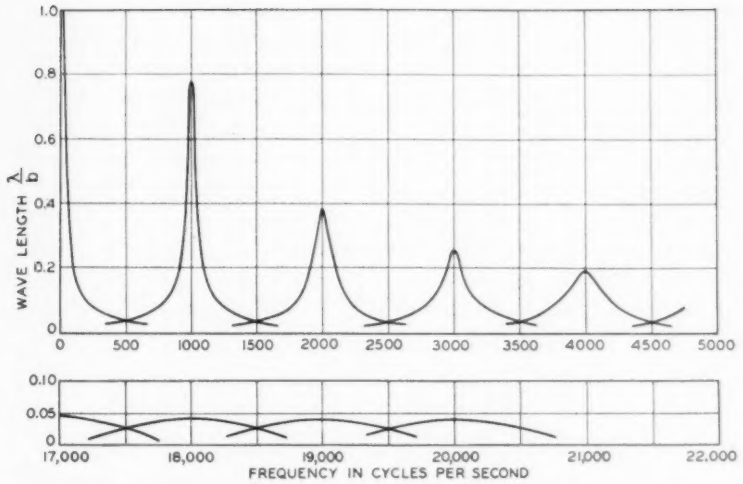


Fig. 8—Correlation between wavelength and frequency of signal components.

In telephotography the frequency of line scanning is usually low and the groups of lines in the frequency spectrum are so closely spaced that such fine grained details of the signal are of little practical importance as far as the electrical parts of the system are concerned. In television, however, these bands are widely spaced, of the order of 1000 cycles or so apart, and such details of the signal are quite important.

As a specific example, it is interesting to plot the frequency spectrum of the television signal that results from scanning a circular area of uniform brightness on a black background. So far as the present theory extends, this may be done by converting the field components of equation (9) into current components with the aid of equations (10) and (11). Taking $b/a = 1.28$, the radius of the circle as $b/3$, and assuming that the field is scanned in 50 lines 20 times per second, we obtain the amplitude-frequency spectrum shown in Fig. 9. Since it is not convenient to show the individual current components—only 20 cycles apart—the curve shows simply the envelope of the peaks of these components. At low frequencies, the energy is largely confined to bands at multiples of the line scanning frequency, 1000 cycles, and to an additional band extending up from zero frequency. In the regions between the bands, the signal components are so small that they do not show when plotted to the same scale. At higher frequencies the signal energy as thus far computed is not confined to such bands. It

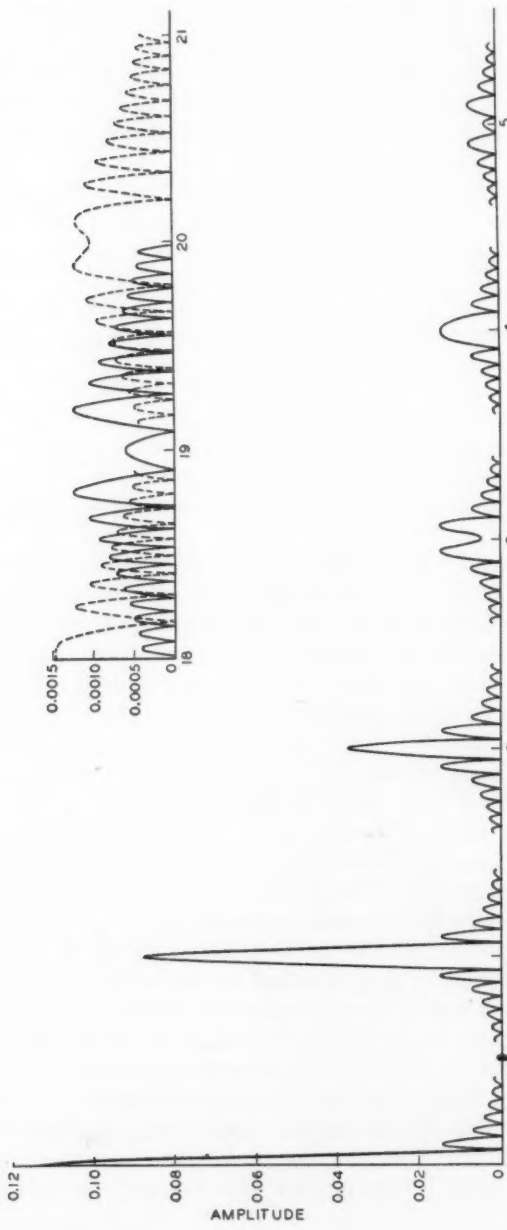


Fig. 9—Amplitudes of signal components for a circular area of brightness (diameter, 1/3 of field length).

will be shown farther on, however, that the effect of the use of a finite aperture for scanning is to confine the signal energy more rigorously to such bands throughout the frequency range.

The theoretical energy distribution for the circular area is in excellent agreement with actual frequency analyses of television currents, which show the energy confined to bands at multiples of the line scanning frequency with apparently empty regions between. It is evident from the theory so far, however, that these regions are not really empty but are filled with weak signal components representing fine details of the subjects; and subjects of greater pictorial complexity than a simple circular area may be devised to give large signal components in such regions. We must therefore look for other factors to explain why these frequency regions do not transmit any appreciable details of an image.

Confusion in the Signal

With the usual method of scanning, one such factor is the confusion of components in the signal. This confusion arises from the fact that two or more image components sloping across the field in different directions may intercept the line of scanning with their crests spaced exactly the same distance apart along this line of scanning. As the scanning point passes over them they thus give rise to signal current components of exactly the same frequency. Consequently the two image components are represented by a single, confused, signal current component that can transmit no information whatever in regard to their relative amplitudes and phases. This confusion evidently depends on the scanning path.

If the image field is scanned in N lines, the velocity v of the scanning point parallel to the y axis is

$$v = \frac{b}{Na} u \tag{12}$$

and the signal frequencies from equation (11) are

$$f = \frac{u}{2a}\left(m + \frac{n}{N}\right). \tag{13}$$

Field components with indices m, n and m', n' such that

$$m + \frac{n}{N} = m' + \frac{n'}{N} \tag{14}$$

give rise to current components of the same frequency.

In other words, the bands of components in the frequency spectrum really overlap. Consequently the components of one band may coincide in frequency with the components of adjacent bands. Such coinciding components are illustrated schematically in Fig. 10.

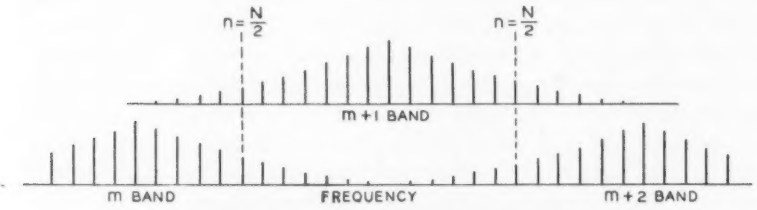


Fig. 10—Coinciding lines of confused bands.

It is obvious that a single a-c. component cannot transmit the separate amplitudes and phases of two or more image components. Consequently the receiving apparatus has no information to judge how the components in the original image are supposed to be distributed in the reproduction.

The situation is most serious where the intensities of coinciding components have the same order of magnitude, that is, at the centers of the frequency regions intermediate to the strong bands. The confusion in these regions is the most important factor that renders them incapable of transmitting any appreciably useful image detail.

On first consideration it would appear that the overlapping of bands in the signal might result in a hopeless confusion. The situation is saved, however, by the fact that components with large n numbers will tend to be weak due to the convergence of the Fourier series, and are further reduced, as will be shown later, by the effects of a scanning aperture of finite size. They therefore do not usually seriously interfere with the stronger components. The interference usually manifests itself in the form of serrations on diagonal lines and occasional moiré effects in the received picture.

Confusion in the signal may be practically eliminated by using an aperture of such a nature that it cuts off all components with n numbers greater than $N/2$, that is, cuts off each band before it reaches the center of the intervening frequency regions so that adjacent bands do not overlap. The practical possibilities of this arrangement will be discussed further below.

The mere elimination of confusion in the signal itself does not necessarily prevent the appearance of extraneous components in the reproduced image. The receiving apparatus itself must be so designed that

when it reproduces all the image components represented by a given signal component, it suitably suppresses all those but the dominant one desired.

Effect of a Finite Aperture at the Transmitting Station

In the preceding pages the scanning aperture has been assumed as infinitesimal in size, or merely a point. In any actual scanning system the necessary finite size of the aperture introduces effects which will now be considered for the transmitting end.

Let us first review briefly the usual theory of this effect when the picture is analyzed simply as a one-dimensional Fourier series. According to equation (3) above, this series is

$$E_1(x) = \sum_{n=-\infty}^{+\infty} A_n \exp i\pi(nx/L).$$

Let ξ be a coordinate fixed with respect to the scanning aperture as shown in Fig. 11 and let the optical transmission of the aperture for

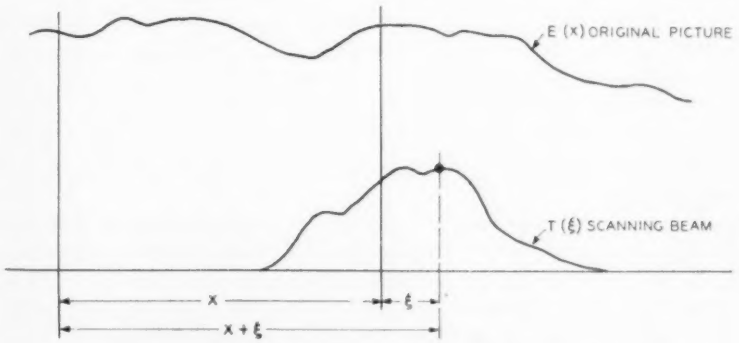


Fig. 11—Analysis of one-dimensional scanning operation.

any value of ξ be $T(\xi)$.[3] Then if x is taken as a coordinate of the origin of ξ the illumination at any point ξ of the aperture is

$$E_1(x + \xi) = \sum_{n=-\infty}^{+\infty} A_n \exp i\pi(n[x + \xi]/L), \qquad (15)$$

[3] This optical transmission may represent either the transparency of an aperture of constant width or the width of an aperture which is a shaped hole in an opaque screen.

so that the total flow of light through the aperture at any position x is

$$F_1(x) = \int_{\text{aperture}^4} T(\xi)E_1(x + \xi)d\xi. \tag{16}$$

Since x is a constant with respect to the integration the exponential term may be factored and the part involving x only may be brought outside the integral sign. This gives

$$F_1(x) = \sum_{n=-\infty}^{+\infty} Y(n)A_n \exp i\pi(nx/L), \tag{17}$$

where

$$Y(n) = \int_{\text{aperture}} T(\xi) \exp i\pi(n\xi/L)d\xi. \tag{17'}$$

For a symmetrical aperture (that is, about the origin of ξ)

$$Y(n) = \int_{\text{aperture}} T(\xi) \cos(\pi n\xi/L)d\xi \tag{17''}$$

and $Y(n)$ in this case is, therefore, a pure real quantity.

The important conclusion to be drawn from equation (17) as to the effect of a finite transmitting aperture is that it multiplies the complex amplitude A_n of each original image component by a quantity $Y(n)$ which is independent of the picture being scanned. This is entirely similar to the effect of a linear electrical network in a circuit, and the quantity $Y(n)$ is quite analogous to the transfer admittance of that network.

The quantity $Y(n)$ has been plotted for variously shaped apertures in Fig. 12. For convenience in comparison, the ordinates of each curve have been multiplied by a numerical factor to make $Y(0) = 1$. The curves show the characteristics that are by this time familiar, which are that the effect of the finite size of the scanning aperture in the transmitter is similar to that of introducing a low-pass filter in the circuit, namely, cutting down the amplitudes of the signal components for which n is numerically high, i.e., the high-frequency components.

The curves are remarkable, however, in that in the useful frequency band (i.e. from $n = 0$ to something like half of the first root of $Y(n) = 0$) all the distributions considered give practically the same transfer admittance if the dimensions of the beam along the direction of scanning are suitably chosen, as has been done in the figure. This results from

[4] The integral is mathematically taken from $-\infty$ to $+\infty$ but the regions outside the aperture give no contribution since the integrand is there equal to zero.

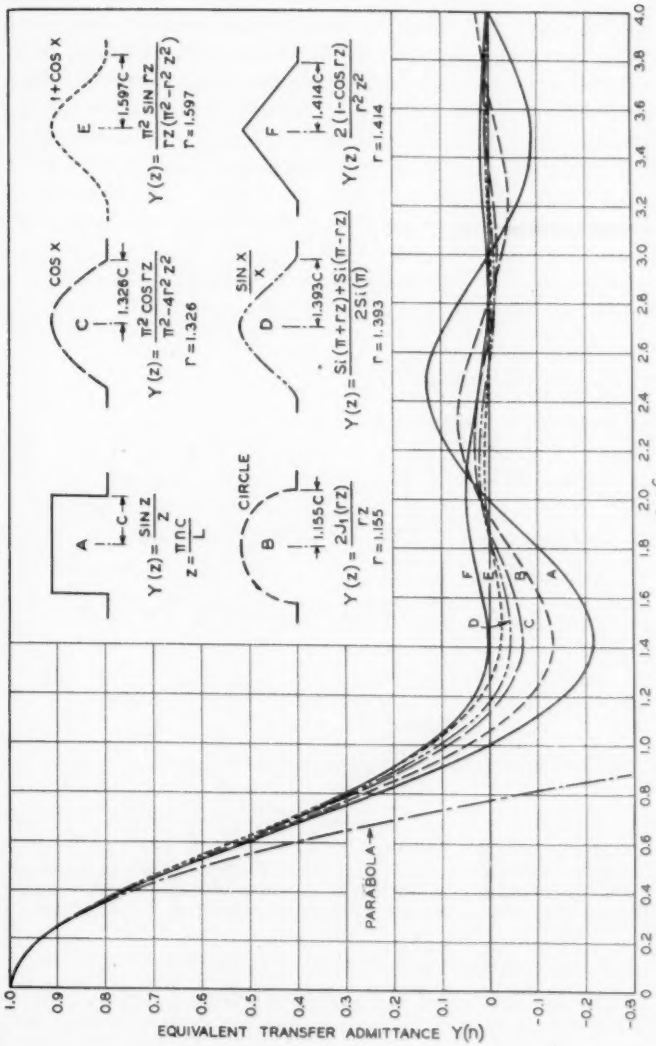


Fig. 12—Equivalent transfer admittances for various distributions of light in the scanning beam.

the physical limitation that the illumination in any part of the beam must be positive, that is, the illumination from one part of the beam must always add to that from another part and cannot subtract from it.[5] This observation enables one to define the resolution of two apertures of different shapes as being equal along a certain direction when their transfer admittances in the useful frequency range show the same filtering effect, if that direction is used as the direction of scanning. This will occur when the radii of gyration (about a normal axis in the plane of each aperture) are equal. According to this definition all the apertures illustrated in Fig. 12 have the same resolution along a horizontal direction.

When the picture is analyzed as a two-dimensional Fourier series the equations which have been given above become

$$E_1(x, y) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} A_{mn} \exp i\pi \left(\frac{mx}{a} + \frac{ny}{b} \right),$$

$$E_1(x + \xi, y + \eta) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} A_{mn} \exp i\pi \left(\frac{m[x + \xi]}{a} + \frac{n[y + \eta]}{b} \right), \tag{18}$$

$$F_1(x, y) = \int\int_{\text{aperture}} T_1(\xi, \eta) E_1(x + \xi, y + \eta) d\xi d\eta, \tag{19}$$

$$F_1(x, y) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} Y_1(m, n) A_{mn} \exp i\pi \left(\frac{mx}{a} + \frac{ny}{b} \right), \tag{20}$$

where

$$Y_1(m, n) = \int\int_{\text{aperture}} T_1(\xi, \eta) \exp i\pi \left(\frac{m\xi}{a} + \frac{n\eta}{b} \right) d\xi d\eta. \tag{20'}$$

For an aperture symmetrical about both ξ and η axes

$$Y_1(m, n) = \int\int_{\text{aperture}} T_1(\xi, \eta) \cos \pi \left(\frac{m\xi}{a} + \frac{n\eta}{b} \right) d\xi d\eta. \tag{20''}$$

[5] The shape of the transfer admittance curve near $n = 0$ depends upon the power of n in the first variable term of the Taylor expansion for $Y(n)$ about $n = 0$, and upon the sign of this term. Assuming a symmetrical aperture, the expansion from equation (17″) is

$$Y(n) = \int T d\xi - \frac{\pi^2 n^2}{2!\, L^2} \int \xi^2 T d\xi + \frac{\pi^4 n^4}{4!\, L^4} \int \xi^4 T d\xi - \cdots.$$

Since T is everywhere positive the first variable term is always in n^2 and negative. The shape of the curve near $n = 0$ is, therefore, always a parabola (indicated in Fig. 12), which can be made the same parabola by suitably choosing the two disposable constants in the aperture. Even after departing from this common parabola, the curves maintain the same general shape over a substantial range; for the next variable term is in n^4 and positive, and has the same order of magnitude for all usual types of apertures. Consequently, the curves for these apertures have approximately the same shape over a wide range extending up from $n = 0$. The results are the same for an unsymmetrical aperture, but the reasoning is more involved.

In the two-dimensional case $T(\xi, \eta)$ is defined, for a hole in an opaque screen, as unity throughout the area of the hole, and zero for the screen. Where the aperture is covered with a non-uniform screen T may take on intermediate values.

The transfer admittances have been calculated for a variety of shapes of aperture in Appendix I. It will be noted that for those types of aperture for which T can be separated into two factors, one a function of ξ only and the other a function of η only, namely, for which

$$T_1(\xi, \eta) = T_\xi(\xi) \cdot T_\eta(\eta), \tag{21}$$

then equation (20′) becomes

$$Y_1(m, n) = \int_{\text{aperture}} T_\xi(\xi) \exp(i\pi m\xi/a)d\xi \int_{\text{aperture}} T_\eta(\eta) \exp(i\pi n\eta/a)d\eta$$
$$= Y_\xi(m) \cdot Y_\eta(n) \tag{22}$$

and Y_ξ and Y_η are each one-dimensional integrals of the type illustrated in Fig. 12.

The rectangular aperture is a simple case of this type. Assume the field to be scanned in N lines and take the dimensions of the aperture, $2c$ and $2d$ parallel to the x and y axes, respectively, as

$$\frac{c}{a} = \frac{d}{b} = \frac{1}{N}. \tag{23}$$

Then

$$Y_\xi(m) = \frac{\sin \pi mc/a}{\pi mc/a}$$

and

$$Y_\eta(n) = \frac{\sin \pi nd/b}{\pi nd/b}$$

and the frequency corresponding to a given signal component mn is, from equation (11)

$$f = \frac{u}{2a}\left(m + \frac{n}{N}\right).$$

Thus, $Y_1(m, n)$ considered as a function of the signaling frequency corresponding to each component of indices mn, consists of a succession of similar curves $u/2a$ cycles apart, corresponding to the successive integral values of m (these curves are themselves really not continuous but consist of a succession of points $u/(2aN)$ cycles apart. For convenience, however, the drawings will always show the curves as con-

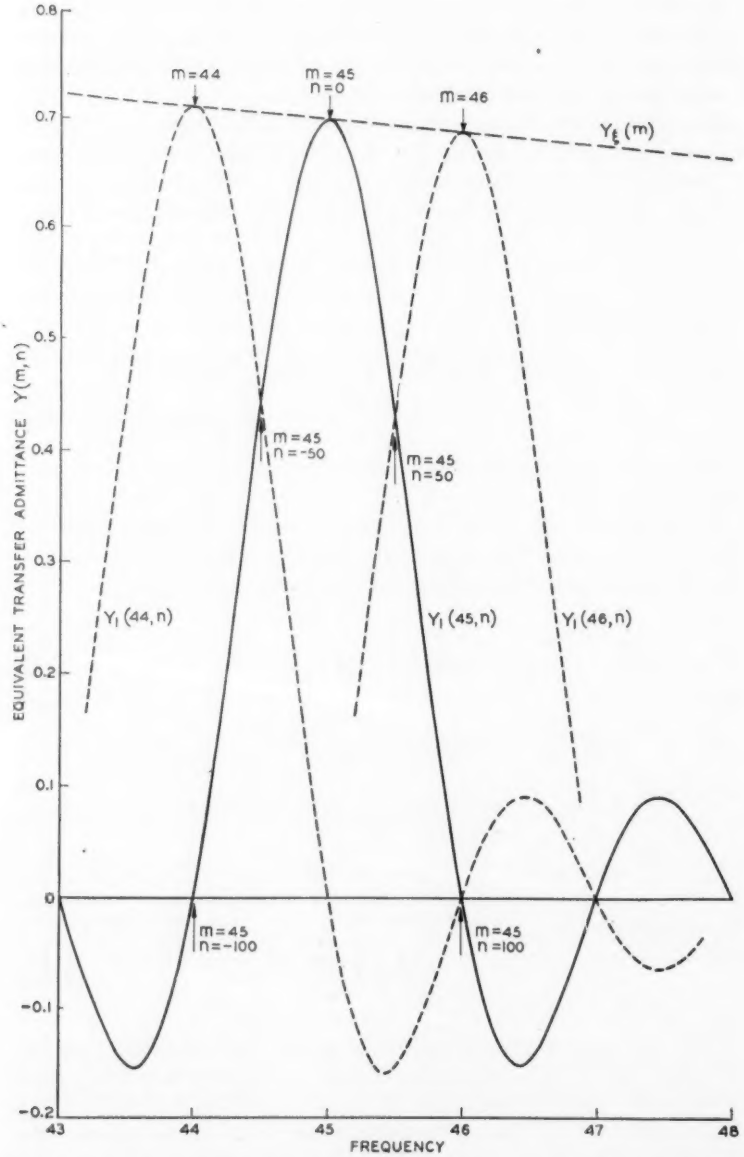


Fig. 13a—Detail of equivalent transfer admittance of aperture for two-dimensional scanning.

tinuous). Each of the curves is of the equation

$$Y_1(m, n) = Y_\xi(m) \frac{\sin \frac{2\pi Nc}{u}\left(f - \frac{mu}{2a}\right)}{\frac{2\pi Nc}{u}\left(f - \frac{mu}{2a}\right)}$$

and therefore has a peak of the value $Y_\xi(m)$ at the point where $n = 0$ or $f = mu/2a$, and trails off from the peak in each direction according to a curve of the same shape as curve "A" in Fig. 12. The successive curves are all of identical shape, but each one is to a reduced scale of ordinates as compared with the preceding (in the useful frequency range) as imposed by the factor $Y_\xi(m)$.

The peaks, it will be noted, occur at the frequencies occupied by what have been called the fundamental components (as distinguished from the satellite lines) in the discussion above on the frequency spectrum of the signal.

Assuming N to be 100 and for simplicity taking the factor $u/2a$ as equal to 1, a plot is shown in Fig. 13a of $Y_1(m, n)$ over a very limited region near the upper end of the useful frequency range. The curve shown in a solid line represents $Y_1(m, n)$ for $m = 45$, and the dotted curves on either side represent the function for $m = 44$ and 46, respectively.

The function has been redrawn for the complete useful range of frequencies and a little beyond, in Fig. 13b, with the frequencies to a logarithmic scale. This logarithmic plot opens out the scale at the low frequencies and enables the fine structure of the function to be indicated there, and still enables the complete range of useful frequencies to be shown without requiring a prohibitive size of drawing (it has, however, the disadvantages that the distortion in the frequency scale then masks the symmetry of the individual curves around the fundamental lines, the similarity of shape of these individual curves, and also the constant frequency separation between the successive fundamental lines).

The function $Y_1(m, n)$, as is clear from equation (22) and Figs. 13a and b, consists of a sort of envelope function $Y_\xi(m)$, "modulated" by a fine structure function $Y_\eta(n)$. The latter function has the value unity at the positions of the fundamental lines in the frequency spectrum of Fig. 7 and diminishes for the satellite lines in the same way that the envelope function diminishes for the fundamental lines away from zero frequency. It will be seen that the envelope function is the only one obtained by the simple one-dimensional analysis. The

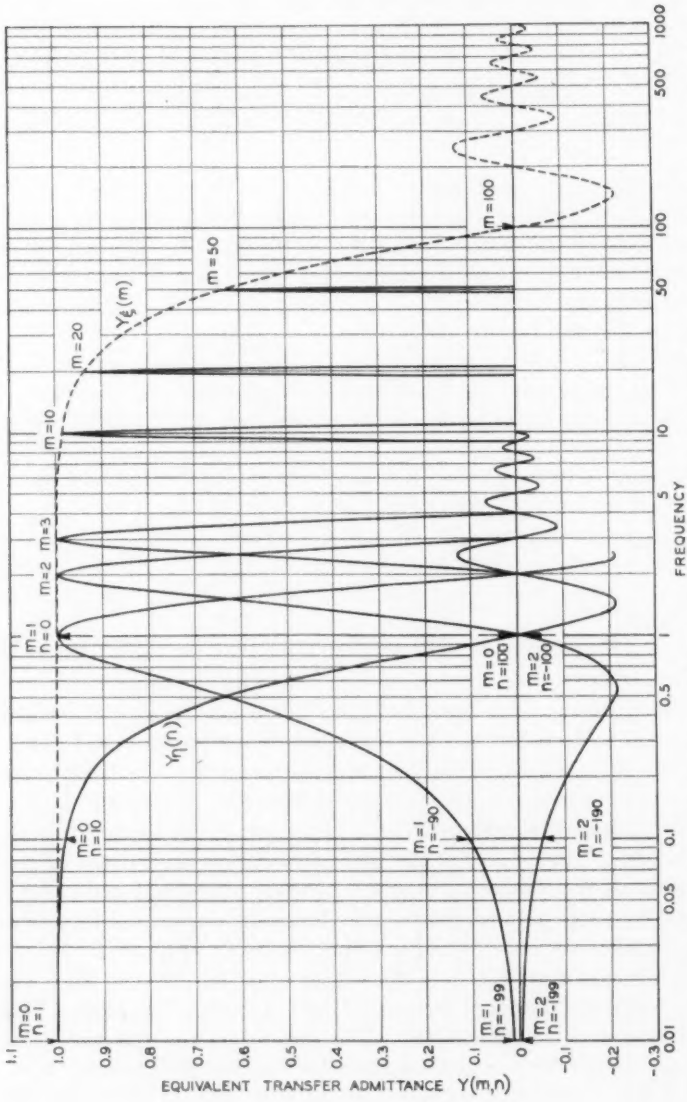


Fig. 13b—Equivalent transfer admittance of aperture for two-dimensional scanning, to a logarithmic frequency scale.

$$E_2(x,y) = f\left(B_{m'n'}\right)$$

where

$$B_{m'n'} = f\left(E_2(x,y)\right)$$

complete function shows, by the very small transfer admittance in the regions half-way between the fundamental lines, an additional reason why the signal currents in these regions will be weak and relatively incapable of transmitting appreciable image detail.

Examination of the other apertures for which computations are given in Appendix I will show that, in general, for all ordinary apertures the same broad phenomena are observed as for the rectangular aperture, although it is not always possible to express the complete function in the simple product form above, in which case the curves for the successive values of m will vary gradually in shape.

The final signal current is proportional to the light flux through the aperture, given in equation (20). Neglecting constant factors it may, therefore, be written as

$$I(t) = \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} Y_1(m, n) A_{mn} \exp i\pi \left(\frac{mu}{a} + \frac{nv}{b} \right) t. \qquad (24)$$

Reconstruction of the Image at the Receiving Station

At the receiving station the signal current is translated back into light to illuminate an aperture moving in synchronism with the one at the sending end. Neglecting constant factors the flow of light $F_2(t)$ to the receiving aperture is

$$F_2(t) = I(t). \qquad (25)$$

Let $E_2(x, y)$ be the resulting apparent illumination (integrated with respect to time) at a point x, y of the reproduced image, or, in telephotography, the integrated exposure of the recording film at this point. This illumination may be expressed as a double Fourier series, similar to equation (7) (but primed subscripts will be used to distinguish them from those of that equation).

$$E_2(x, y) = \sum_{m'=-\infty}^{+\infty} \sum_{n'=-\infty}^{+\infty} B_{m'n'} \exp i\pi \left(\frac{m'x}{a} + \frac{n'y}{b} \right), \qquad (26)$$

where

$$B_{m'n'} = \frac{1}{4ab} \int_{-b}^{+b} \int_{-a}^{+a} E_2(x, y) \exp -i\pi \left(\frac{m'x}{a} + \frac{n'y}{b} \right) dx dy. \qquad (27)$$

Reproduction of detail in the image may be studied by comparing these components with the corresponding ones of the original image.

The apparent illumination is the same as if the aperture traced a single strip across repeated fields in the xy plane as illustrated in Fig. 14, and all of the repeated fields included between $y = -b$ and

$y = +b$ were cut out and superposed to form the image. Let $E_3(x, y)$ be the illumination of this strip. Then, since the exponential

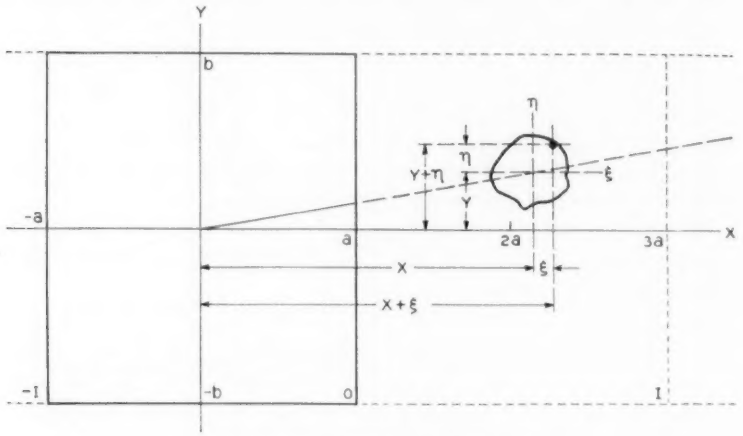


Fig. 14—Analysis of received picture.

factor of the integrand in equation (27) is periodic in x and identically reproduced in each of the fields, the integral is equal to

$$B_{m'n'} = \frac{1}{4ab}\int_{-b}^{+b}\int_{-\infty}^{+\infty} E_3(x, y) \exp -i\pi\left(\frac{m'x}{a} + \frac{n'y}{b}\right) dxdy. \quad (28)$$

The limits $-b$ to $+b$ in y and the infinite limits in x may be used because the illumination is zero everywhere outside of the strip.

Again taking a coordinate system $\xi\eta$ fixed with respect to the aperture, such that

$$x = \xi + ut, \qquad y = \eta + vt \quad (29)$$

the instantaneous illumination of any point covered by the aperture is, neglecting constant factors,

$$T_2(\xi, \eta)I(t) = T_2(x - ut, y - vt)I(t). \quad (30)$$

The total illumination of any point xy in the image strip is thus

$$E_3(x, y) = \int_{-\infty}^{+\infty} T_2(x - ut, y - vt)I(t)dt.$$

Substitution in integral (28) and a change in the order of integration

gives

$$B_{m'n'} = \frac{1}{4ab}\int_{-\infty}^{+\infty}\int_{-b}^{+b}\int_{-\infty}^{+\infty} T_2(x - ut, y - vt)I(t) \cdot \exp -i\pi\left(\frac{m'x}{a} + \frac{n'y}{b}\right) dxdydt. \quad (31)$$

Changing to the $\xi\eta$ system

$$B_{m'n'} = \frac{1}{4ab}\int_{-\infty}^{+\infty}\int_{-b-\eta t}^{b-\eta t}\int_{-\infty}^{+\infty} T_2(\xi, \eta)I(t) \exp -i\pi\left(\frac{m'u}{a} + \frac{n'v}{b}\right) t \cdot \exp -i\pi\left(\frac{m'\xi}{a} + \frac{n'\eta}{b}\right) d\xi d\eta dt. \quad (32)$$

This integral may be considered as the surface integral of a function $\varphi(\eta, t)$ taken over a strip shaped area shown in Fig. 15, in elements of

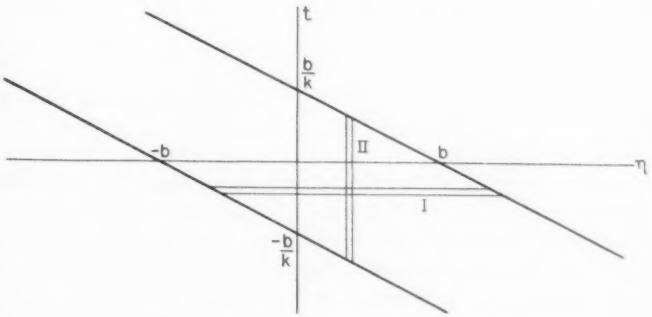


Fig. 15—Equivalent integration regions.

the type indicated as *I*. From this it may be seen that, when also integrated in elements of the type indicated as *II*,

$$\int_{-\infty}^{+\infty}\int_{-b-\eta t}^{b-\eta t} \varphi(\eta, t)d\eta dt = \int_{-\infty}^{+\infty}\int_{(-b-\eta)/v}^{(b-\eta)/v} \varphi(\eta, t)dtd\eta. \quad (33)$$

Consequently

$$B_{m'n'} = \frac{1}{4ab}\int_{-\infty}^{+\infty}\int_{(-b-\eta)/v}^{(b-\eta)/v}\int_{-\infty}^{+\infty} T_2(\xi, \eta)I(t) \exp -i\pi\left(\frac{m'u}{a} + \frac{n'v}{b}\right) t \cdot \exp -i\pi\left(\frac{m'\xi}{a} + \frac{n'\eta}{b}\right) d\xi dtd\eta. \quad (34)$$

Consider now the intensity $B_{m'n'}$ of a final reproduced picture component m', n' resulting from a single component m, n in the signal as

expressed by equation (24). The integral becomes

$$B_{m'n'} = \frac{A_{mn}Y_1(m,n)}{4ab}\int_{-\infty}^{+\infty}\int_{(-b-\eta)/v}^{(b-\eta)/v}\int_{-\infty}^{+\infty} T_2(\xi,\eta) \cdot \exp i\pi\left(\frac{m-m'}{a}u+\frac{n-n'}{b}v\right)t \cdot \exp -i\pi\left(\frac{m'\xi}{a}+\frac{n'\eta}{b}\right)d\xi dt d\eta. \quad (35)$$

It will be noted that the exponential function of t is periodic in t, one of the periods being $t_0 = 2b/v$. Furthermore, this is just the difference between the upper and lower limits in t. Hence the integral in t may be written

$$I = \int_0^{t_0} \exp i\pi\left(\frac{m-m'}{a}u+\frac{n-n'}{b}v\right)t dt.$$

This integral is zero except when

$$\frac{m-m'}{a}u+\frac{n-n'}{b}v = 0, \quad (36)$$

in which case

$$I = t_0. \quad (37)$$

The meaning of these last few equations is clear. It is, as would be expected, that a signal component m, n does not give rise to all components m', n' in the final received picture, but that these latter components are in general zero unless m' and n' satisfy a definite relationship with m and n, expressed by equation (36). A somewhat unexpected result is, however, that equation (36) allows some other m', n' components besides the normal one for which $m' = m$ and $n' = n$. That is to say, a given signal component m, n in the line will reproduce in the final picture not only a corresponding m, n component, but as has been foreshadowed in the discussion on confusion in the signal, it will also reproduce certain other components with different indices.

Let us consider first, however, the reproduction of the normal component for which $m' = m$ and $n' = n$, which is obviously allowed by equation (36). The amplitude B_{mn} is then, neglecting constant factors,[6]

$$B_{mn} = A_{mn}Y_1(m,n)\,Y_2(m,n), \quad (38)$$

where

$$Y_2(m,n) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} T_2(\xi,\eta)\exp -i\pi\left(\frac{m\xi}{a}+\frac{n\eta}{b}\right)d\xi d\eta. \quad (38')$$

[6] The constant factor neglected as compared with equation (35) is $t_0/(4ab)$. The t_0 is the period of image repetitions (or "frame period"). It appears here because the brightness of a single image depends on how quickly it is reproduced.

The quantity Y_2 it will be noted is almost the same, for the receiving aperture, as the Y_1 is in equation (20′) for the sending aperture. Thus, on the normally reproduced component the receiving aperture merely adds whatever filtering action it has to that which has already been caused by the sending aperture.

As noted, in addition to this normal component, the integral (35) exists in general for other values of m' and n' and thus gives rise to extraneous components in the reproduced image. If equation (36) is applied particularly to the usual system of scanning in N lines in which as in equation (12), $v = ub/(Na)$, it becomes

$$m + \frac{n}{N} = m' + \frac{n'}{N}. \tag{39}$$

For values of m' and n' satisfying equation (39), the reproduced component has the complex amplitude (neglecting constant real factors)

$$B_{m'n'} = A_{mn} Y_1(m, n)\, Y_2(m', n'). \tag{40}$$

Looking back at equation (14) and comparing it with equation (39) it may be seen that these components correspond in indices to the original image components that are confused in the signal to give only one signal component. The result is, therefore, after all quite reasonable from a physical point of view. For when a signal of a certain frequency is transmitted over the line the receiving apparatus has no information by which to judge which component in the original picture it is supposed to represent. So, as shown by equation (40) it impartially reproduces every one of the components it could possibly represent, each component with the intensity and phase it would have if it were really the one intended to be represented by the signal. The components are then all superimposed in the picture.

From this development it is clear that the process of scanning an image field in strips and reproducing it in a similar manner not only reproduces the components of the original image but also introduces extraneous components. The reproduced field thus consists of two superposed fields: a normal image built up from the normally reproduced components, and an additional field of extraneous components. Although not really independent, it is convenient to consider these two fields as existing separately, and thus to think of the normal image field as having an extraneous field superposed on it.

Considering the normal field alone, we may term the reproduction of its detail as the *reproduction of* NORMAL *detail*. There is a loss in such reproduction, for both the transmitting and receiving apertures intro-

duce a relative loss in the reproduction of the shorter wave components. Consequently there is a loss of definition in the finer grained details of the normal image. This type of distortion due to aperture loss may be termed *simple omission of detail.*

In addition to the simple omission of detail, the normal image is masked by the presence of the extraneous field. The more pronounced features of this field are the line structure and serrated edges that it superposes on the normal image. Its presence is not only displeasing, but it also masks the normal image components and thus results in a further loss of useful detail. This type of loss may be termed a *masking of detail* or a *masking loss.* It is true that the extraneous components may sometimes give rise to an illusory increase in resolution across the direction of scanning in special cases where they add on to the diminished normal components in just the right phase and magnitude to bring the latter back to their phases and intensities in the original image, giving no resultant distortion whatever. (In all such cases, however, to obtain this benefit it is necessary to effect a quite accurate register between the original image and the scanning lines or the distortion is very large. Such accurate registering is generally impractical and may be definitely impossible if the registry required for one portion of the image conflicts with that required in another portion. Such cases may, therefore, in general be disregarded.)

The Reproduction of Normal Detail

The preceding theory permits a numerical calculation of the reproduction of detail in the normal image. This is given directly by equation (38) above.

In order to make some of the discussion in the following pages more concrete and specific the sending and receiving apertures will be taken alike; this condition, therefore, gives $[Y(m, n)]^2$ as a measure of how well the various components are reproduced. If a picture be assumed in which all the original components have the same amplitude then $[Y(m, n)]^2$ is the amplitude of the reproduced normal components.

The relative admittance for any given pair of apertures may be calculated from equations (20′) or (38′). Such calculations have been made for various apertures and the results summarized in Appendix II.

The admittance of an aperture is not in general uniquely determined by the wave length of a component, but also depends on the orientation of the component with respect to the aperture. The admittances of reasonably shaped apertures do, however, decrease in general with increasing numerical values of the indices m and n; and the shorter wave components are, therefore, in general, less faithfully reproduced than the longer wave ones.

A circular aperture furnishes a simple example of such reproduction—because its admittance, from its symmetrical shape, is a unique function of the wave length of a component. In other words a circular aperture reproduces normal detail equally well in all directions. We may, therefore, simply plot $[Y(m, n)]^2$ as a function of the component wave length as in Fig. 16, and this single curve is a measure of how

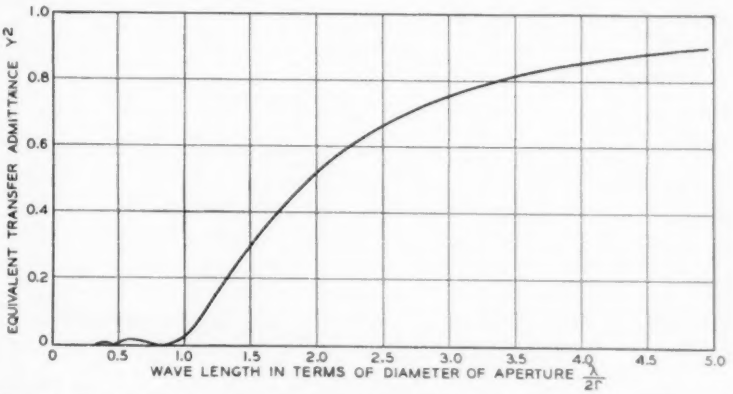


Fig. 16—Equivalent transfer admittance for circular apertures at both sending and receiving ends, *vs.* wavelength.

well the various normal components are reproduced. The shorter wave components are practically omitted in the reproduction of an image.

Other apertures do not reproduce normal image detail equally well in all directions because their admittances depend on the slope of a component. To simplify the consideration of such apertures we may resort to a practice commonly used in discussing telephotographic or television systems, and that is, we may take the resolution along the direction of scanning and across the direction of scanning separately as criteria of their performance.

Neglecting the small slope of scanning lines with respect to the x axis of the image field, the admittance of an aperture for components normal to the direction of scanning is $Y(m, 0)$. Consequently, we may take $[Y(m, 0)]^2$ as a measure of the reproduction of normal detail along the direction of scanning. In a similar manner we may take $[Y(0, n)]^2$ as a measure of the reproduction of normal detail across the direction of scanning.

It thus follows that an aperture gives the same resolution of normal detail along the direction of scanning and across the direction of scan-

ning when the two admittances $Y(m, 0)$ and $Y(0, n)$ are substantially equal for components of the same wave length over the useful range. Circular apertures, square apertures and other apertures that are suitably symmetrical fulfill this condition exactly, and consequently give equal resolution of normal detail in the two directions.

The curve $[Y(0, n)]^2$ has been plotted, by way of illustration for a rectangular aperture, in the middle line of Fig. 17.

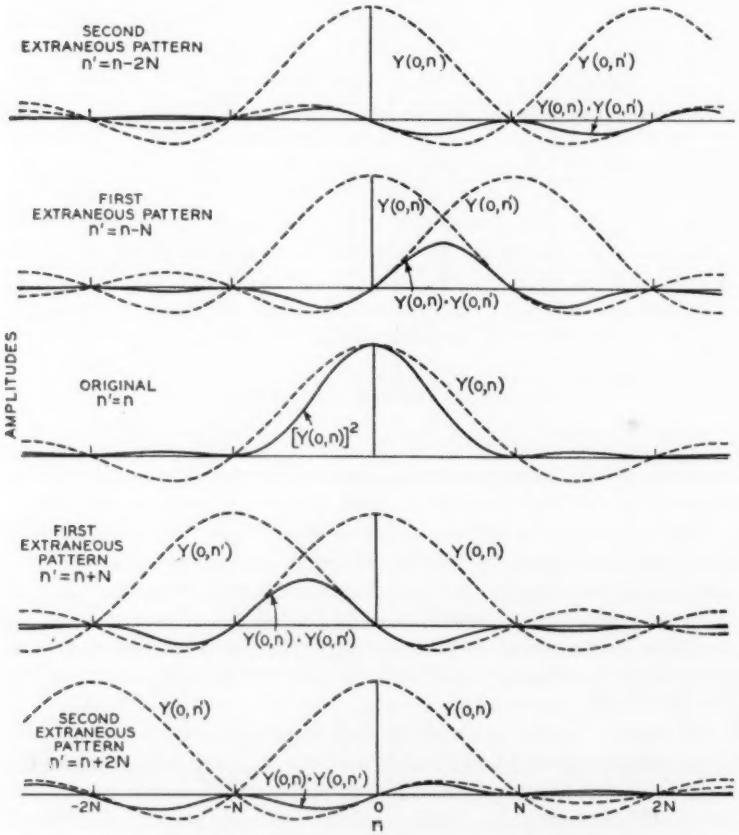


Fig. 17—Reproduction of original and extraneous patterns.

It may be noted incidentally that the simple omission of detail which occurs in the reproduction of normal components is quite similar to the loss of resolution that an image suffers when it is reproduced through

an imperfect optical system. Specifically the effect of a sending or receiving circular aperture alone, or $Y(m, n)$, is the same as that caused by an optical system which reproduces a mathematical point in the original as a circle of uniform illumination (circle of confusion) in the image of the same size (with respect to the image) as the scanning aperture. The effect of the two apertures in tandem, or $[Y(m, n)]^2$, may be very closely simulated by a circle of confusion of about twice the area of either aperture, as can be judged from the discussion which has been given above regarding the curves in Fig. 12.

The Extraneous Components

It will be clearly understood that the discussion immediately preceding has been confined entirely to the normal image components, that is, to the image that would be seen if no extraneous components were present. In particular, it should be clear that the reproduction of normal detail equally well in the direction of scanning and across the direction of scanning does not mean that the details of the total resultant image will be seen equally well in the two directions, for the extraneous components will to a certain extent mask the normal image.

In the same manner as for the normally reproduced components, the amplitudes of the extraneous components, according to the preceding theory, are given by equation (40) above, where

$$m' = m + \mu,$$
$$n' = n - \mu N,$$

where

$$\mu = \text{an integer} = m' - m = (1/N)(n - n'). \tag{41}$$

The composite transfer admittance $Y(m, n) \cdot Y(m', n')$ may therefore be taken as a measure of the extent to which the extraneous components are introduced. If a picture be assumed in which all the original components have the same amplitude then $Y(m, n) \cdot Y(m', n')$ is the amplitude of the extraneous components.

A given original component of indices m, n gives rise to a whole series of extraneous components, m', n', as μ ranges from 1 up through the positive integers and -1 down through the negative integers. As an illustration we have plotted the case of a rectangular aperture of a width just equal to the scanning pitch, in Fig. 17, which has just been referred to in considering the normal components. The two lines marked "first extraneous pattern" show the relative amplitudes for μ equal to 1 and -1, respectively, and those marked "second extran-

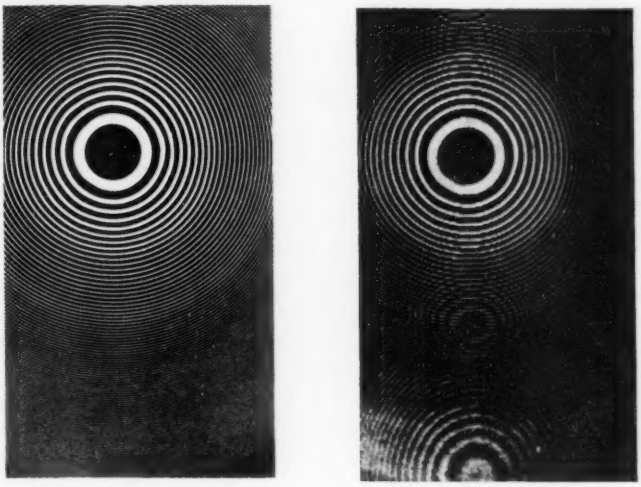
eous pattern," for μ equal to 2 and -2, respectively, for $m = 0$. (The shift from $m' = 0$ to $m' = \pm 1$ and ± 2 has been ignored since if N is at all large this has a negligible effect on $Y(m', n')$, as may be noted from Fig. 13.) An examination of Fig. 17 and a consideration of the nature of $Y(m, n)$ and $Y(m', n')$ shows that the principal interference effect will come from the pattern for which $|\mu| = 1$, and that the relative amplitudes become very small as μ increases in absolute magnitude. In general, therefore, only the first extraneous pattern may be considered as of really serious importance. Considering this pattern in Fig. 17 it will be seen that the amplitude $Y(0, n)\ Y(0, n')$ increases as $|n'|$ increases from zero, the extraneous components becoming more and more comparable to the normal components. At $N/2$, both components are of the same amplitude, and the extraneous components are therefore masking the normal components. It will be noted that the index region at $N/2$ corresponds to the centers of the relatively empty regions in the frequency spectrum of the signal. The large masking effect caused by the extraneous components explains why such small signal energy as exists in these regions is almost completely incapable of transmitting any useful image detail.

It will be noted that the components with values of $|n'|$ in the neighborhood of $N/2$ and greater are in general almost parallel to the direction of scanning. The masking loss will therefore be greatest across the direction of scanning and practically negligible along the direction of scanning. This is quite reasonable because the extraneous components constitute the line structure of the reproduced image, and should therefore cause the greatest loss of detail across the direction of scanning.

For clarity in the explanation up to this point, masking loss has been discussed as if an extraneous component could only mask the normal component with which its indices happened to coincide. In reality the masking is of a more serious nature. An extraneous component undoubtedly obscures any normal component that has about the same wave length and the same slope across the field even though it does not exactly coincide in these characteristics.

More detailed curves than Fig. 17, showing the amplitudes of the extraneous components have been prepared in Appendix II. These also show the results for other index values of m than zero, and for other than the simple rectangular aperture. The results indicate that the extraneous patterns diminish in intensity progressively as more overlap is tolerated between adjacent scanning lines, at the expense, of course, of increased aperture loss for the normal components. This point will be taken up again below.

The reality of these extraneous components is strikingly demonstrated in Fig. 18, for which we are indebted to Mr. E. F. Kingsbury.



a. Original. *b*. Transmitted.

Fig. 18—Fresnel zone plate.

This shows at (*a*) the original of a Fresnel zone plate and at (*b*) the picture after transmission through a telephotographic system. The first extraneous pattern is very prominent in the lower corner of (*b*) and a detailed study of the slope and spacing of the extraneous striations shows them to be in exact accord with the theory which has been given.[7]

The special case of the extraneous components which are formed when the original consists of a flat field is of some interest due to the high visibility of these components under such a condition. This scanning line structure is quite familiar as an imperfection in many pictures

[7] The extraneous pattern, although it is (and should be according to the theory) very nearly a transposed reproduction of the original pattern, must not be confused with a long delayed echo of that original pattern. In other words, if only the lower half instead of the whole of (*a*) had been transmitted, the lower half of (*b*) would still have been exactly as it is, the extraneous components being generated entirely irrespective of whether components representing a similar configuration exist in other portions of the original or not.

In the region about half-way between the centers of the normal picture and the first extraneous picture the resulting pattern gives very much the appearance of another set of extraneous components. It is not such, however, that successive rings are not really bright and dark, as they would be in the case of a genuine extraneous component, but alternating uniform gray and striped black and white, so that the average intensity along the circumference of a ring is independent of the diameter of the ring, except for some photographic non-linearity.

transmitted by telephotography and television. It can be removed only by insuring that $Y(m', n')$ shall vanish whenever $n' = N$, so that

$$Y(0, 0) \cdot Y(\mu, \mu N) = 0.$$

The requirement can be met for the elementary shapes of apertures A, E and F of Fig. 12, but cannot be met in the others. In these other cases the overlap between adjacent scanning lines is usually adjusted so that the requirement is met for $\mu = \pm 1$, to remove the most serious pattern. Thus, for example, for the circular aperture B this requires an overlap of around 25 per cent.

The Reproduction of Detail

In optical instruments the reproduction of detail is usually measured by what is called the "resolving power" which in turn is defined from the smallest separation between two mathematical point (or parallel line) sources of light in the original which can be distinguished as double in the reproduced image.

For the present it is perhaps simpler to consider another criterion of the resolving power, namely, the shortest element length in an image resembling a telegraph signal, used as an original, which can be recognizably reproduced with certainty in the received picture. For reasons that have already been mentioned above it is necessary to insist that the received picture be recognizable with certainty without any registry requirement between the original image and the scanning lines.

Using this criterion for the resolution along the direction of scanning and assuming the apertures at the sending and receiving ends to be rectangular and of the same length with respect to the picture size, the minimum signal element required for a recognizable picture (as set by the apertures as distinguished from the electrical transmission circuits) will be of about the length of either aperture. For other shapes of aperture the minimum element length will be very nearly the length of the equivalent rectangular aperture using the term "equivalent" in the same sense that it was used in the discussion regarding Fig. 12.

According to the same criterion, for the resolution across the direction of scanning the minimum element length required for recognizable transmission, in the case of a rectangular aperture of width equal to the scanning pitch, will be twice the scanning pitch. It will be noted that this is twice the length which would be required if only the normal image components were reproduced, and this difference may be considered as a measure of the degradation caused by the masking effect of the extraneous components for this arrangement of apertures.

This figure for the degradation must be taken with a certain reserve, partly because the exact telegraph theory for the criterion of resolution considered has really been inferred rather than presented in complete logical form, and partly because the figure may be expected to vary according to the criterion of resolution chosen. Some rough studies have indicated the degradation to be materially less if the more conventional criterion of resolution (two parallel line sources of light) were used.

This degradation may be estimated in another manner. In Appendix II the extraneous components have been computed for a variety of apertures and degrees of overlap between adjacent scanning lines. In Fig. 19 there have been plotted the maximum amplitudes of these

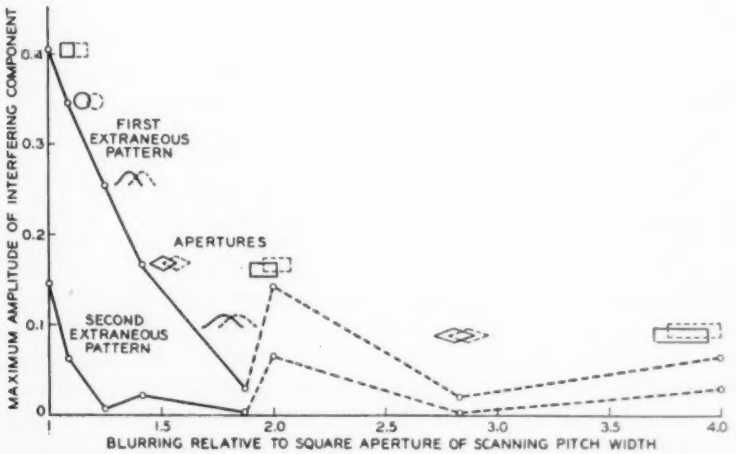


Fig. 19—Magnitude of extraneous components as a function of resolution.

extraneous components in each case (the first and second extraneous patterns being plotted separately) as a function of the relative coarseness of resolution for the normal image alone. This latter quantity is taken relative to a rectangular aperture of width equal to the scanning pitch, and, for example, for a rectangular aperture of width equal to twice the scanning pitch, is represented by the figure 2. For convenience, above the various points have been inserted small diagrammatic representations of the corresponding apertures. Also for convenience the points have been arbitrarily connected together.

From inspection of Fig. 19 several conclusions may be drawn, namely,

1. Considering apertures of a given shape, the more overlap allowed between adjacent scanning lines the weaker will be the extraneous patterns but the coarser will be the reproducible detail in the normal image.

2. Not all shapes of aperture are equally efficient in suppressing extraneous components, and at the same time retaining a given resolution of normal detail. Of the shapes considered, the rectangular aperture is least efficient in this respect, and the full-wave sinusoidal aperture (E in Fig. 12), is the most efficient.

3. Although not proved, it may be inferred from the figure that the finest resolution in the normal image that can be obtained (assuming a given scanning pitch) without showing a first order extraneous pattern on a flat field, is that obtained with the rectangular aperture of width equal to the scanning pitch.

4. With the most suitable aperture it is possible practically to suppress the extraneous components, at the expense of coarsening the normal reproducible detail to slightly under twice that given by the rectangular aperture just mentioned.

The last point in particular enables us to draw a conclusion in regard to the degradation contributed by the extraneous components. For a rectangular aperture of width equal to the scanning pitch it appears that the degradation amounts to a little less than doubling the coarseness of resolution to normal detail. This substantially checks the estimate which has already been made above. It may further be surmised for all the other shapes of aperture shown with a value of abscissa under 2 that as the degradation contributed by the extraneous components is reduced, the coarseness of resolution to normal detail is increased to just about make up for this, and that in the overall picture the minimum element length which can be recognizably reproduced remains substantially constant at about twice the scanning pitch.[8] For aperture arrangements with values of abscissa over 2, either the inefficiency in suppressing extraneous components, or the unnecessarily large overlap, tends to coarsen the overall resolution to a minimum elementary length greater than twice the scanning pitch. In this region the line connecting the points has been dotted.

[8] It may very well be that even if all these aperture arrangements transmit an about equal amount of information they do not give the same psychological satisfaction to the viewer at the receiving end. The general effect of a square aperture of scanning pitch width is to give a "snappy" appearance, disturbed, however, by the presence of the extraneous patterns. When these are removed, keeping the overall resolution about the same, the appearance becomes "woolly" or "fuzzy."

An Estimate of the Idle Frequency Regions

As mentioned at the beginning of this paper, the frequency regions between the strong bands appear to be empty when examined with a frequency analyzer of limited level range, or when a narrow band elimination filter is used in connection with visual observations of the reproduced image. These regions are not really completely empty, but do contain weak signal components as shown by the preceding theory, which are not, however, particularly useful inasmuch as, in the final result, they give rise about equally to components simulating the original picture and to masking extraneous components. The regions may, therefore, be considered as idle.

The factors determining the extent of these idle regions are too complicated to permit an exact theoretical evaluation of their width, but an estimate may be attempted from an inspection of Fig. 12 and of the curves given in Appendix II.

From Fig. 12 and the experience that along the direction of scanning the minimum recognizably transmitted elementary signal length is the length of a rectangular aperture it can be deduced that in the absence of extraneous components the useful band of an aperture extends up to the point where its relative admittance, for a single aperture, is in the neighborhood of 0.65. For two apertures in tandem the corresponding relative admittance is $0.65^2 = 0.42$.

Now in Fig. 27 of Appendix II the extraneous components are very small and may be considered negligible. According to the above criterion, therefore, the useful frequency band constitutes approximately 54 per cent of the total space. The idle frequency regions would, therefore, occupy the remainder, or 46 per cent of the total space.

Experimental examination of a television signal with a narrow band elimination filter gave the width of the idle regions as 50 to 60 per cent of the total space. This was for a field scanned with a circular aperture giving a one-quarter overlap of scanning strips. The discrepancy for a quantity so vaguely defined is not large but is probably due to incomplete utilization of even the theoretically active region by the television set because of inherent imperfections in parts of the complete system outside the scanning mechanism proper.

The width of the individual idle bands is then about half the frequency of repetition of scanning lines. For most systems of telephotography this runs in the order of magnitude of one cycle per second, making the waste regions very narrow and close together. For systems of television the waste bands come in much more significant "slices," although the same fraction of the frequency space is wasted. For ex-

ample, in a 50-line system the waste bands are each about 500 cycles wide. In a system using a single sideband of one million cycles width the waste bands are each about 3300 cycles wide.

These idle frequency regions naturally lead to the questions whether (*a*) there is any way of segregating all of the relatively useless signal components in one region of the frequency spectrum so that the useful parts of the signal may be transmitted over a channel of about half the width, or (*b*) whether it would be worth while placing other communication channels in these waste regions. It must be realized, however, that even when the complete frequency space is utilized (by any one of a number of possible schemes), the required frequency band for transmitting a picture of given detail at a given rate is still only halved as compared with the simple system considered above, which is not a change in order of magnitude. The problems of transmitting the wide band of frequencies necessary, for example, in television, while lessened, therefore still remain.

APPENDIX I

The calculation of $Y_1(m, n)$ according to equation (20′) is, for the three simple apertures here considered, a straightforward mathematical process which will therefore not be reproduced. The results are plotted in the form of charts in the conventional manner for functions of two variables, namely as a series of contours, one of the two variables being kept constant for each contour. This constant value changes progressively for each successive contour.

The variables are taken as m and n, multiplied by parameters depending on the sizes of the scanning aperture and of the picture. Because of the obvious symmetry of the function, only half of each chart has been drawn. In order to avoid confusion the contours have been dotted when $|m|$ is greater than the first root of $Y_1(m, 0) = 0$. In one case the contour is shown in a dashed line when $|m|$ is equal to this root. Constant factors in the scale of ordinates have been neglected, to make $Y_1(0, 0) = 1$.

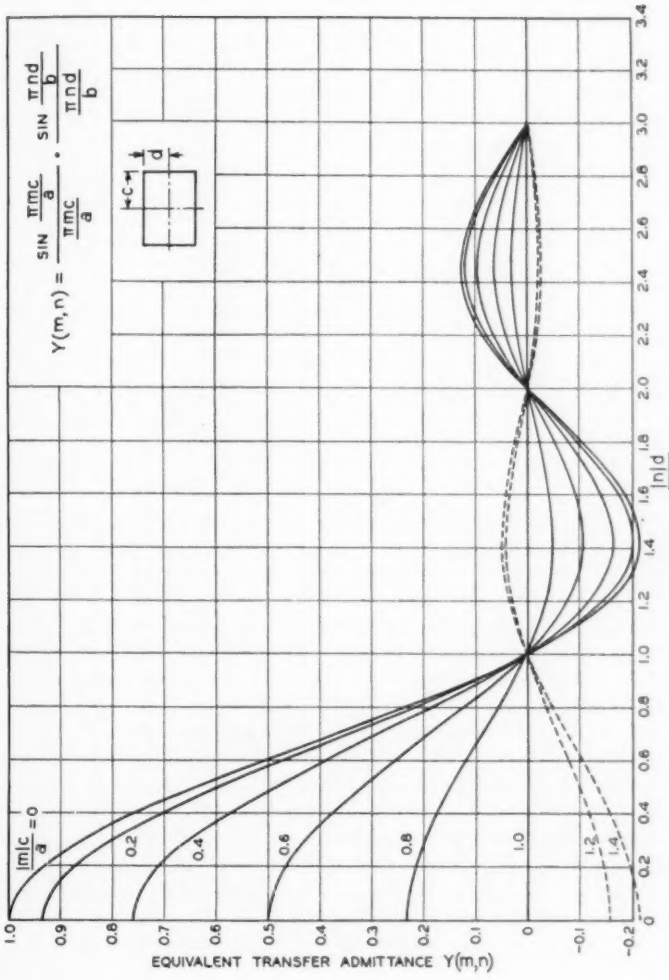


Fig. 20—Rectangular aperture.

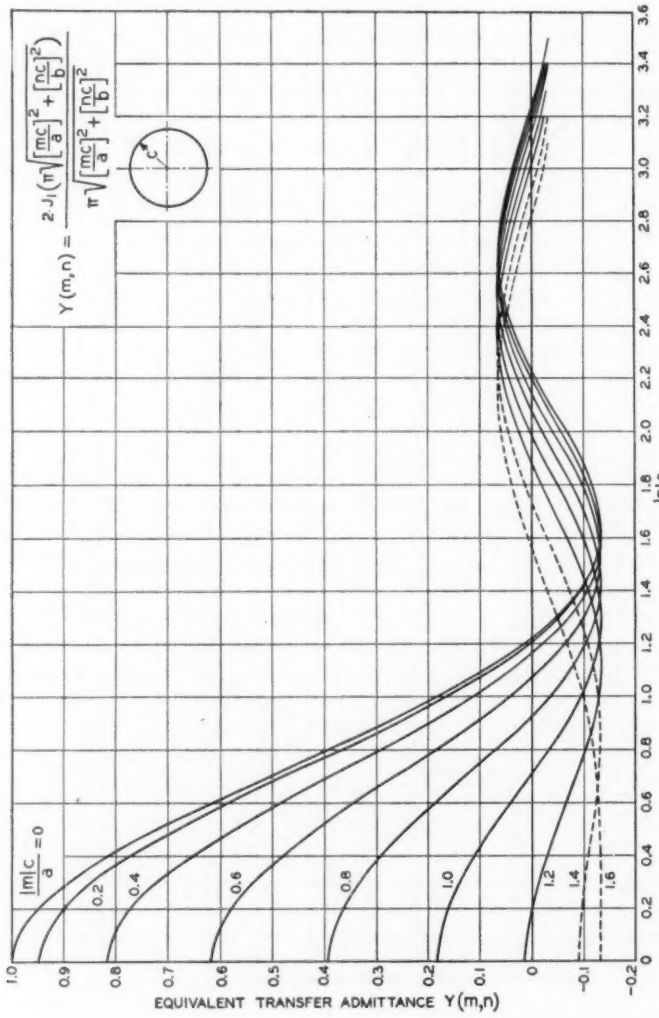


Fig. 21—Circular aperture.

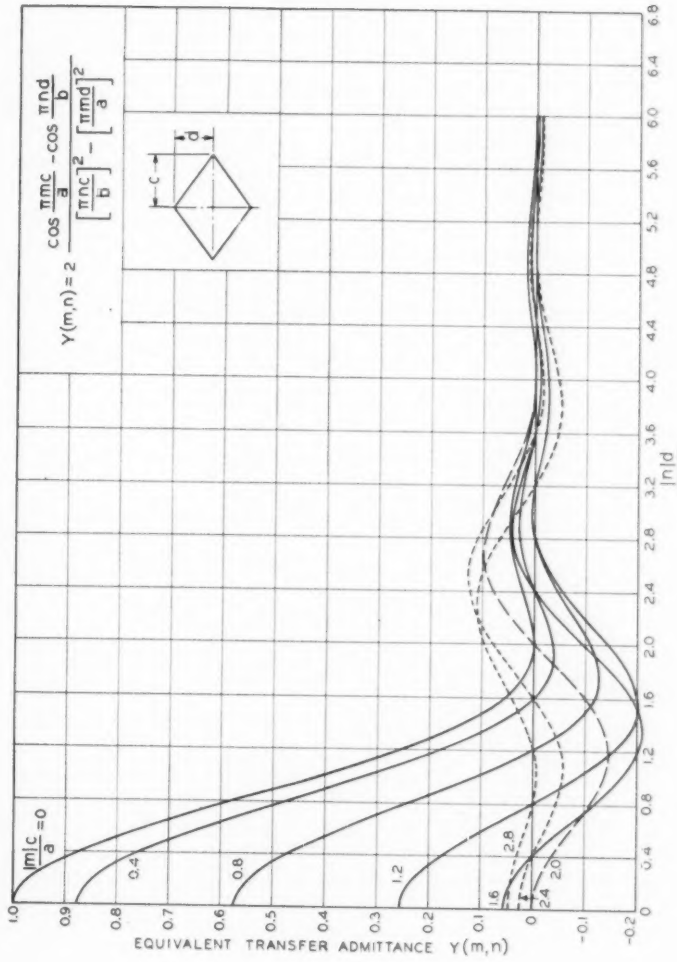


Fig. 22—Diamond shaped aperture.

APPENDIX II

As in the case of Appendix I, the calculation of $Y_1(m, n) \cdot Y_2(m', n')$ is a straightforward mathematical procedure which will not be reproduced. The results are again presented in the form of charts.

In these charts the intensities of the principal extraneous components are indicated by solid lines, while the higher order extraneous components are indicated by dotted lines. The normal components have been indicated by dashed lines.

It should be explained that what has really been plotted is $Y_1(m, n) \cdot Y_2(m, n')$ rather than $Y_1(m, n) \cdot Y_2(m', n')$. This is because the difference between m' and m, when multiplied by the parameters chosen for the charts, varies with the proportions of the scanning system used. As discussed in the text with regard to Fig. 17, however, this difference between m' and m has a negligible effect for any system employing a useful number of scanning lines.

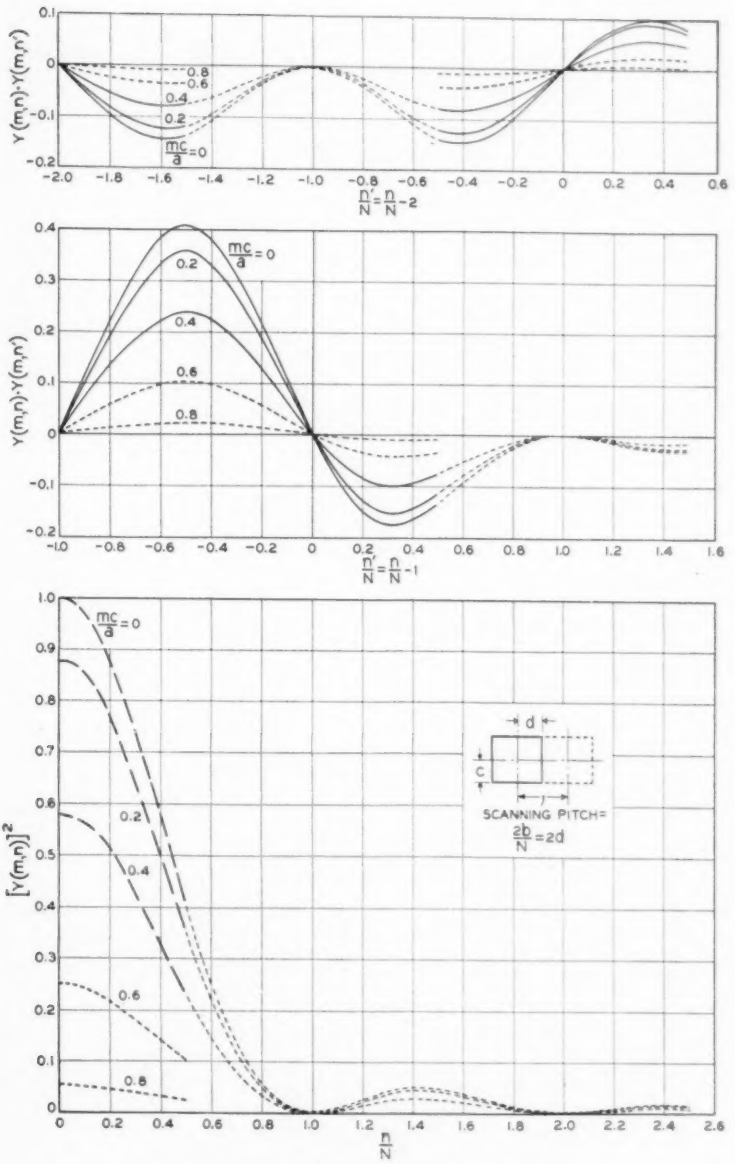


Fig. 23—Rectangular aperture with no overlap.

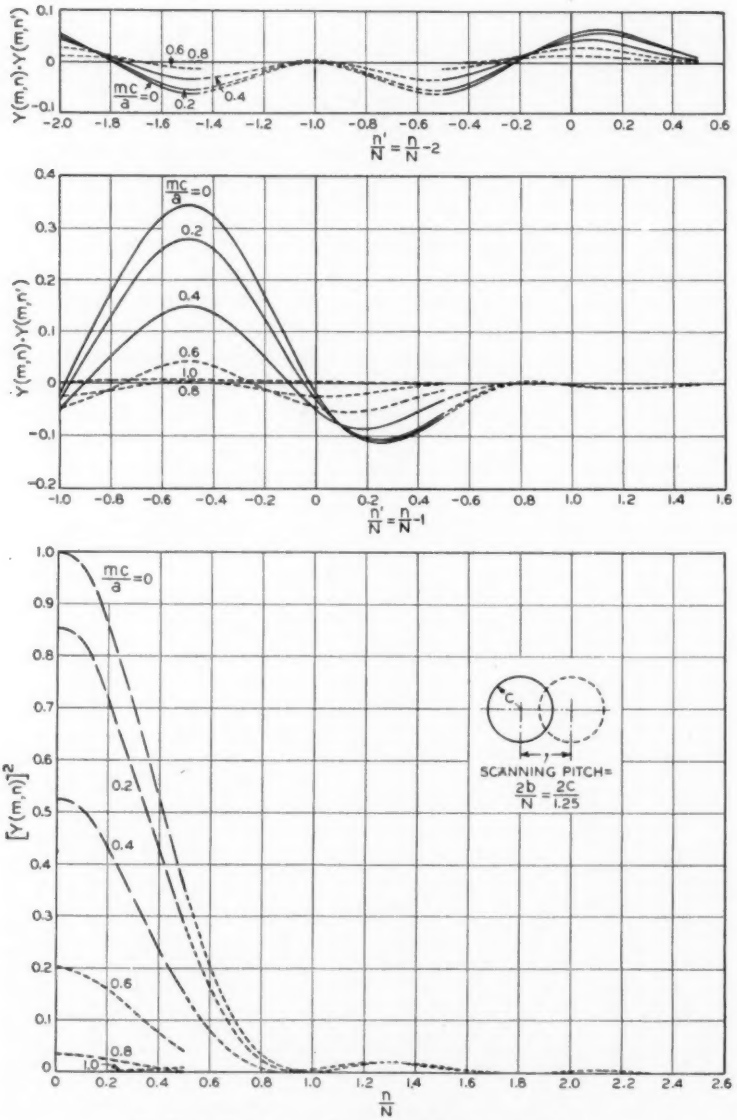


Fig. 24—Circular aperture with 25 per cent overlap.

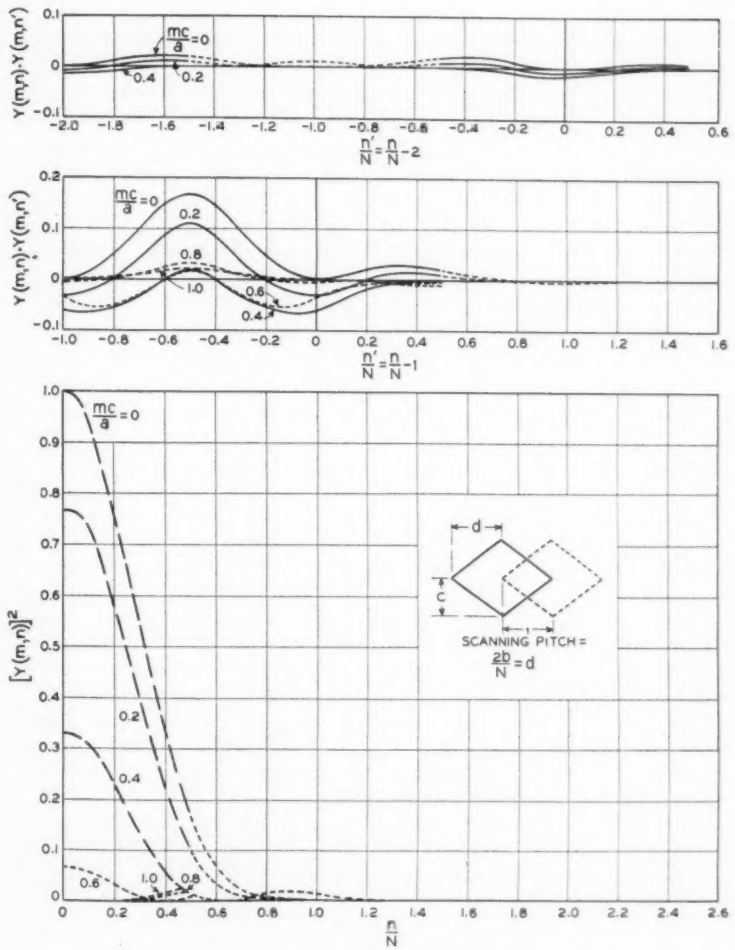


Fig. 25—Diamond shaped aperture with half diagonal overlap.

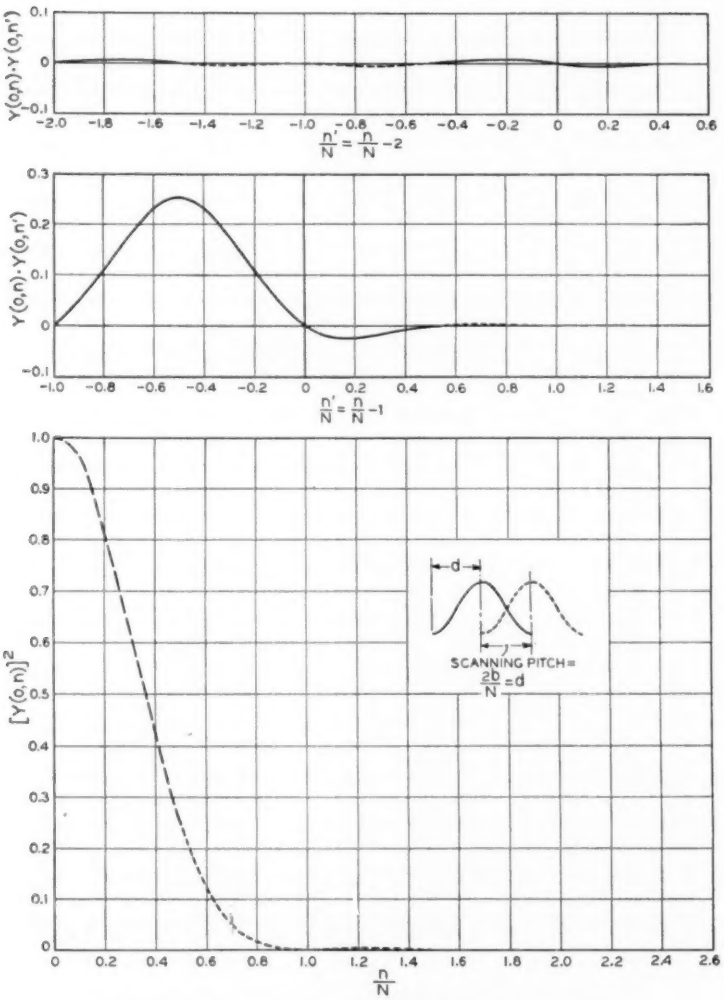


Fig. 26—Sinusoidal aperture with half wavelength overlap.

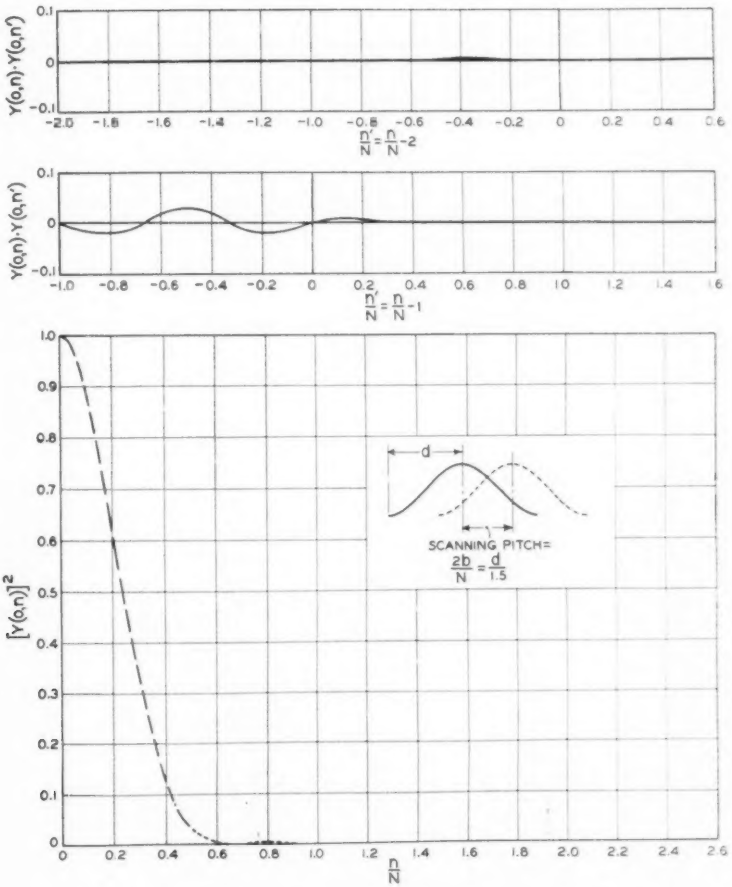


Fig. 27—Sinusoidal aperture with two-thirds wavelength overlap.

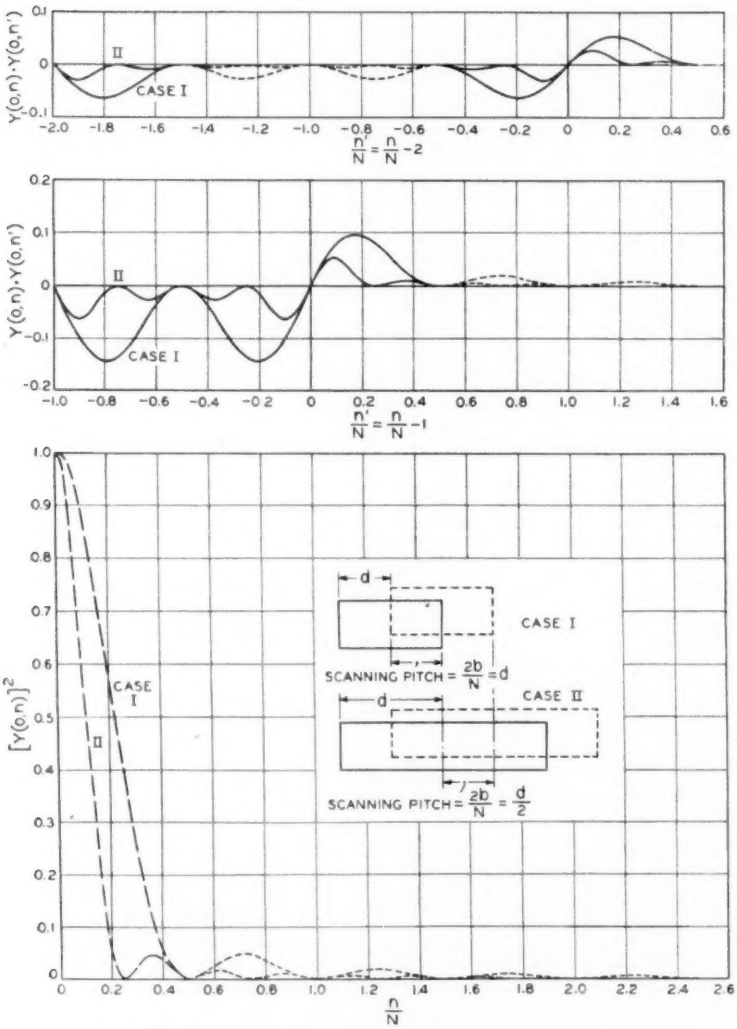


Fig. 28—Rectangular aperture with different degrees of overlap.

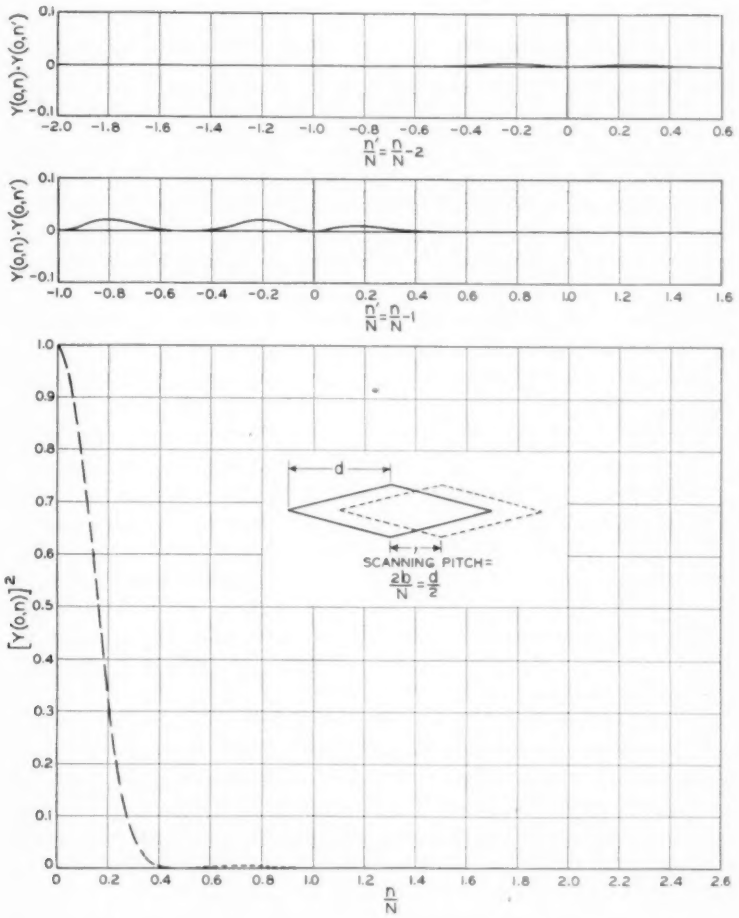


Fig. 29—Diamond shaped aperture with three-quarters diagonal overlap.

Abstracts of Technical Articles from Bell System Sources

The Thermionic Work Function and the Slope and Intercept of Richardson Plots.[1] J. A. BECKER and W. H. BRATTAIN. This article is a critical correlation of the slope and intercept of experimental Richardson lines with the quantities appearing in *theoretical* equations based on thermodynamic and statistical reasoning. The equation for *experimental* Richardson lines is $\log i - 2 \log T = \log A - b/2.3\,T$; A and b are constants characteristic of the surface, i is the electron emission current in amp./cm.^{2}, T is the temperature in degrees K, $\log A$ is the intercept and $-b/2.3$ is the slope of experimental lines. Statistical *theory* based on the Fermi-Dirac distribution of electron velocities in the metal shows that i should be given by $\log i - 2 \log T = \log U(1 - r) - w/2.3\,T$, where U is a universal constant which has the value 120 amp./cm.^{2} °K^2, r is the reflection coefficient, and w is the *work function.* A correlation of the experimental and theoretical equations shows that $b = w - T dw/dt$, and $\log A = \log U(1 - r) - (1/2.3) dw/dT$. Only when r is 0 and the work function is independent of the temperature, is it correct to say that the slope is $-w/2.3$ and that the intercept has the universal value $\log U$. But even when w is a function of T, it follows from a thermodynamic argument that the slope is given by $-h/2.3$, where the heat function h is defined by $h = (L_p/R) - (5/2)T$, L_p is the heat of vaporization per mol at constant pressure. The heat function is related to the work function by the equation $h = w - T dw/dT$.

From experimental and theoretical arguments it is deduced that the reflection coefficient is probably negligibly small. Hence we conclude that *for most surfaces the work function varies with temperature,* since the intercepts of Richardson lines are rarely equal to log 120. This conclusion is to be expected since on Sommerfeld's theory, w depends on the number of free electrons or atoms per cm.^{3}, which in turn varies with temperature due to thermal expansion.

The photoelectric work function should equal the thermionic work function but should not in general be equal to -2.3 times the slope of the Richardson line. The Volta potential between two surfaces having work functions w_1 and w_2 should equal $(w_1 - w_2)k/e$ rather than $2.3k/e$ times the difference between the slopes of the Richardson lines for the two surfaces. The data from photoelectric and Volta potential meas-

[1] *Phys. Rev.*, May 15, 1934.

urements support the conclusion that the work function depends on temperature.

Fundamental Concepts in the Theory of Probability.[2] THORNTON C. FRY. Three commonly accepted definitions of the word "probability" are discussed critically, with regard both to logical soundness and to practical utility. Two major theses are presented: first, that each definition has utilitarian merits which render it especially valuable within its own field; second, that the objection of logical redundancy which is so frequently raised against the Laplacian definition can equally well be raised against the other two definitions.

Wide-Range Recording.[3] F. L. HOPPER. The recent improvements in sound quality resulting from the extension of the frequency and intensity ranges are the results of coordinated activity in recording equipment and processes, reproducing equipment, and theater acoustics. This paper discusses the recording phase of the process. A wide-range recording channel consists essentially of the moving-coil microphone, suitable amplifiers, a new recording lens, and certain electrical networks.

The characteristics of such a system, from the microphone to and including the processed film, are shown. Other factors fundamentally associated with wide-range recording, such as monitoring, film processing, the selection of takes in the review room, and re-recording, are also discussed. The changes brought about by this system of recording result, first, in a greater freedom of expression and action on the part of the actor; and, second, in a much greater degree of naturalness and fidelity than has been previously achieved.

Iron Shielding for Telephone Cables.[4] H. R. MOORE. Voltages of fundamental and harmonic frequencies, induced along communication cables by neighboring power or electric railway systems, can be reduced by the electromagnetic shielding action of the sheath, if this is grounded continuously or at the ends of the exposure. The shielding, particularly at the fundamental frequency, is improved greatly by the provision of a steel tape armor, while a surrounding iron pipe conduit effects a very great improvement at both the fundamental frequency and the higher harmonics.

This paper presents methods for the quantitative prediction of the shielding, expressed by a "shield factor" or the fraction to which

[2] *American Mathematical Monthly*, April, 1934.
[3] *Jour. S. M. P. E.*, April, 1934.
[4] *Electrical Engineering*, February, 1934.

a disturbing voltage is reduced. Necessary impedance data are given for numerous iron-surrounded cable constructions and working charts are supplied for the convenient determination of the shielding obtainable with commercially available steel tape armored cables.

On the basis of data presented in this paper, prediction of the shielding to be obtained from steel tape armored cable sheaths or those inclosed in iron pipes is concluded to be both feasible and practical. With internal impedances measurable on short length samples of a chosen construction, the accuracy of prediction is limited principally by the precision to which the disturbing field and the grounding resistances of the cable sheath may be determined. Either of the constructions discussed is capable of effecting a high order of shielding against low frequency induction and practically complete protection from harmonic disturbances. Field observations on installed cables, both tape armored and in pipe conduit, have verified the computational methods presented.

Propagation of High-Frequency Currents in Ground Return Circuits.[5] W. H. WISE. The electric field parallel to a ground return circuit is calculated without assuming that the frequency is so low that polarization currents in the ground may be neglected. It is found that the polarization currents may be included by replacing the r in Carson's well-known formulas by $r\sqrt{li(\epsilon - 1)/2c\lambda\sigma}$. The problem to be solved is that of calculating the electric field parallel to an alternating current flowing in a straight, infinitely long wire placed above and parallel to a plane homogeneous earth. Carson's derivation of this field is based on three restricting assumptions: (1) The ground permeability is unity; (2) the wave is propagated with the velocity of light and without attenuation; (3) the frequency is so low that polarization currents may be neglected. The first of these restrictions is usually of no consequence and the formula would be quite complicated if the permeability were not made unity. As pointed out in a later paper by Carson, the second restriction amounts merely to assuming reasonably efficient transmission. The effect of the third restriction begins to be noticeable at about 60 kilocycles. The object of the present paper is the removal of the third restriction.

Acoustical Requirements for Wide-Range Reproduction of Sound.[6] S. K. WOLF. The extension of the frequency and volume ranges in recording and reproducing sound has aroused a greater and more critical

[5] *Proc. I. R. E.*, April, 1934.
[6] *Jour. S. M. P. E.*, April, 1934.

consciousness of the importance of theater acoustics. It follows that higher fidelity in reproduction excites greater intolerance of the needless distortion caused by poor acoustics of the theater. To cope with the new situation, engineers have developed new instruments for acoustical analysis, which provide greater precision and facility in detecting defects and in determining the necessary corrections.

In addition to instrumental developments there have been concurrent advances in acoustical theory and practice. The result is that the more stringent requirements imposed on the acoustics of the theater by the enlarged frequency and volume ranges can be fulfilled adequately and practically. The paper discusses the requirements and describes some of the available methods for complying with them.

Contributors to this Issue

C. B. AIKEN, B.S., Tulane University, 1923; M.S. in Electrical Communication Engineering, Harvard University, 1924; M.A. in Physics, 1925; Ph.D., 1933. Geophysical research and exploration with Mason, Slichter and Hay, Madison, Wisconsin, 1926–28. Bell Telephone Laboratories, 1928–. Dr. Aiken has been engaged in work on aircraft communication equipment, broadcast receiver design, centralized radio systems and common frequency broadcasting.

A. W. CLEMENT, B.S. in Electrical Engineering, University of Washington, 1925; M.A., Columbia University, 1929. Bell Telephone Laboratories, Apparatus Development Department, 1925–. Mr. Clement has been engaged in the development of various types of transmission networks, such as electric wave filters and equalizers.

KARL K. DARROW, B.S., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., University of Chicago, 1917. Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Darrow has been engaged largely in writing on various fields of physics and the allied sciences.

I. E. FAIR, B.S., in Electrical Engineering, Iowa State College, 1929. Bell Telephone Laboratories, Radio Research Department, 1929–. Mr. Fair has been engaged in experimentation on piezo-electric crystals for frequency control.

FRANK GRAY, B.S., Purdue, 1911; Ph.D., University of Wisconsin, 1916. Western Electric Company, Engineering Department, 1919–25. Bell Telephone Laboratories, 1925–. Dr. Gray has been engaged in work on electro-optical systems.

H. S. HAMILTON, B.S. in Electrical Engineering, Tufts College, 1916. American Telephone and Telegraph Company, Engineering Department, 1916–18; Department of Development and Research, 1918–34. Bell Telephone Laboratories, 1934–. Mr. Hamilton has been engaged exclusively in toll transmission work, including telephone repeaters, program transmission and carrier telephone systems.

F. R. LACK, B.Sc., Harvard University, 1925; Engineering Department, Western Electric Company, 1913–22; First Lieutenant, Signal Corps, A.E.F., 1917–19; Harvard University, 1922–25. Bell Tele-

phone Laboratories, 1925–. Mr. Lack has been engaged in experimental work connected with radio communication.

W. P. MASON, B.S. in Electrical Engineering, University of Kansas, 1921; M.A., Columbia University, 1924; Ph.D., 1928. Bell Telephone Laboratories, 1921–. Dr. Mason has been engaged in investigations on carrier transmission systems and more recently in work on wave transmission networks, both electrical and mechanical.

R. C. MATHES, B.Sc., University of Minnesota, 1912; E.E., 1913. Western Electric Company, Engineering Department, 1913–25. Bell Telephone Laboratories, 1925–. Mr. Mathes has been concerned with the early history of the repeater development program, the application of vacuum tube amplifiers in a variety of fields, and the application of voice controlled switching circuits in the toll telephone plant. As Associate Wire Transmission Research Director he carries on investigations relating to the transmission of speech over wire systems.

PIERRE MERTZ, A.B., Cornell University, 1918; Ph.D., 1926. American Telephone and Telegraph Company, Department of Development and Research, 1919–23, 1926–34. Bell Telephone Laboratories, 1934–. Dr. Mertz has been engaged in special problems in toll transmission, chiefly in telephotography, television, and cable carrier systems.

G. W. WILLARD, B.A., University of Minnesota, 1924; M.A., 1928; Instructor in Physics, University of Kansas, 1927–28; Student and Assistant, University of Chicago, 1928–30. Bell Telephone Laboratories, 1930–. Mr. Willard's work has had to do with special problems in piezo-electric crystals for frequency control.

S. B. WRIGHT, M.E. in Electrical Engineering, Cornell University, 1919. Engineering Department and Department of Development and Research, American Telephone and Telegraph Company, 1919–34. Bell Telephone Laboratories, 1934–. Mr. Wright is engaged in transmission development work on voice-operated systems and wire connections to radio telephone stations.